Theory and applications of Benford’s law to fraud detection, or: Why the IRS should care about number theory!

Steven J Miller (Brown University)
Mark Nigrini (Saint Michael’s College)

sjmiller@math.brown.edu
http://www.math.brown.edu/~sjmiller

IRS (Boston Offices), March 28th, 2008
Summary

- Review Benford’s Law.
- Discuss examples and applications.
- Sketch proofs.
- Describe open problems.
Caveats!

- Not all fraud can be detected by Benford’s Law.
Caveats!

- Not all fraud can be detected by Benford’s Law.

- A math test indicating fraud is *not* proof of fraud: unlikely events, alternate reasons.
Caveats!

- Not all fraud can be detected by Benford’s Law.
- A math test indicating fraud is *not* proof of fraud: unlikely events, alternate reasons.
Notation

- **Logarithms:** $\log_B x = y$ means $x = B^y$. 
Notation

- **Logarithms**: \( \log_B x = y \) means \( x = B^y \).
  - Example: \( \log_{10} 100 = 2 \) as \( 100 = 10^2 \).
Notation

- **Logarithms:** $\log_B x = y$ means $x = B^y$.
  
  - Example: $\log_{10} 100 = 2$ as $100 = 10^2$.
  
  - $\log_B (uv) = \log_B u + \log_B v$. 
Notation

- **Logarithms**: \( \log_B x = y \) means \( x = B^y \).
  - Example: \( \log_{10} 100 = 2 \) as \( 100 = 10^2 \).
  - \( \log_B(uv) = \log_B u + \log_B v \).
  - \( \log_{10}(100 \cdot 1000) = \log_{10}(100) + \log_{10}(1000) \).
Notation

- **Logarithms**: \( \log_B x = y \) means \( x = B^y \).
  - Example: \( \log_{10} 100 = 2 \) as \( 100 = 10^2 \).
  - \( \log_B (uv) = \log_B u + \log_B v \).
  - \( \log_{10}(100 \cdot 1000) = \log_{10}(100) + \log_{10}(1000) \).

- **Set Theory**:
  - \( \mathbb{Q} = \) rational numbers = \( \{p/q : p, q \text{ integers}\} \).
Notation

- **Logarithms**: $\log_B x = y$ means $x = B^y$.
  - Example: $\log_{10} 100 = 2$ as $100 = 10^2$.
  - $\log_B (uv) = \log_B u + \log_B v$.
  - $\log_{10} (100 \cdot 1000) = \log_{10} (100) + \log_{10} (1000)$.

- **Set Theory**:
  - $\mathbb{Q} =$ rational numbers $= \{p/q : p, q \text{ integers}\}$.
  - $x \in S$ means $x$ belongs to $S$. 
Notation

- **Logarithms:** $\log_B x = y$ means $x = B^y$.
  - Example: $\log_{10} 100 = 2$ as $100 = 10^2$.
  - $\log_B(uv) = \log_B u + \log_B v$.
  - $\log_{10}(100 \cdot 1000) = \log_{10}(100) + \log_{10}(1000)$.

- **Set Theory:**
  - $\mathbb{Q} = \text{rational numbers} = \{p/q : p, q \text{ integers}\}$.
  - $x \in S$ means $x$ belongs to $S$.
  - $[a, b] = \{x : a \leq x \leq b\}$.
Notation

- **Logarithms**: \( \log_B x = y \) means \( x = B^y \).
  - Example: \( \log_{10} 100 = 2 \) as \( 100 = 10^2 \).
  - \( \log_B (uv) = \log_B u + \log_B v \).
  - \( \log_{10}(100 \cdot 1000) = \log_{10}(100) + \log_{10}(1000) \).

- **Set Theory**:
  - \( \mathbb{Q} = \) rational numbers = \( \{ p/q : p, q \text{ integers} \} \).
  - \( x \in S \) means \( x \) belongs to \( S \).
  - \( [a, b] = \{ x : a \leq x \leq b \} \).

- **Modulo 1**:
Notation

- **Logarithms:** \( \log_B x = y \) means \( x = B^y \).
  - Example: \( \log_{10} 100 = 2 \) as \( 100 = 10^2 \).
  - \( \log_B(uv) = \log_B u + \log_B v \).
  - \( \log_{10}(100 \cdot 1000) = \log_{10}(100) + \log_{10}(1000) \).

- **Set Theory:**
  - \( \mathbb{Q} = \) rational numbers = \( \{ p/q : p, q \text{ integers} \} \).
  - \( x \in S \) means \( x \) belongs to \( S \).
  - \( [a, b] = \{ x : a \leq x \leq b \} \).

- **Modulo 1:**
  - Any \( x \) can be written as integer + fraction.
Notation

- **Logarithms:** \( \log_B x = y \) means \( x = B^y \).
  - Example: \( \log_{10} 100 = 2 \) as \( 100 = 10^2 \).
  - \( \log_B (uv) = \log_B u + \log_B v \).
  - \( \log_{10}(100 \cdot 1000) = \log_{10}(100) + \log_{10}(1000) \).

- **Set Theory:**
  - \( \mathbb{Q} = \) rational numbers = \( \{ p/q : p, q \text{ integers} \} \).
  - \( x \in S \) means \( x \) belongs to \( S \).
  - \( [a, b] = \{ x : a \leq x \leq b \} \).

- **Modulo 1:**
  - Any \( x \) can be written as integer + fraction.
  - \( x \mod 1 \) means just the fractional part.
Notation

- **Logarithms:** $\log_B x = y$ means $x = B^y$.
  - Example: $\log_{10} 100 = 2$ as $100 = 10^2$.
  - $\log_B(uv) = \log_B u + \log_B v$.
  - $\log_{10}(100 \cdot 1000) = \log_{10}(100) + \log_{10}(1000)$.

- **Set Theory:**
  - $\mathbb{Q} =$ rational numbers $= \{p/q : p, q$ integers$\}$.
  - $x \in S$ means $x$ belongs to $S$.
  - $[a, b] = \{x : a \leq x \leq b\}$.

- **Modulo 1:**
  - Any $x$ can be written as integer + fraction.
  - $x$ mod 1 means just the fractional part.
  - Example: $\pi$ mod 1 is about .14159.
Benford’s Law: Newcomb (1881), Benford (1938)

### Statement

For many data sets, probability of observing a first digit of \( d \) base \( B \) is \( \log_B \left( \frac{d+1}{d} \right) \); base 10 about 30% are 1s.
Benford’s Law: Newcomb (1881), Benford (1938)

**Statement**

For many data sets, probability of observing a first digit of \(d\) base \(B\) is \(\log_B \left( \frac{d+1}{d} \right)\); base 10 about 30% are 1s.

- Not all data sets satisfy Benford’s Law.
Benford’s Law: Newcomb (1881), Benford (1938)

Statement

For many data sets, probability of observing a first digit of \( d \) base \( B \) is \( \log_B \left( \frac{d+1}{d} \right) \); base 10 about 30% are 1s.

- Not all data sets satisfy Benford’s Law.
- Diamond Long street \([1, L]\): \( L = 199 \) versus \( L = 999 \).
Benford’s Law: Newcomb (1881), Benford (1938)

Statement

For many data sets, probability of observing a first digit of \( d \) base \( B \) is \( \log_B \left( \frac{d+1}{d} \right) \); base 10 about 30% are 1s.

- Not all data sets satisfy Benford’s Law.
  - Long street \([1, L]\): \( L = 199 \) versus \( L = 999 \).
  - Oscillates between \( 1/9 \) and \( 5/9 \) with first digit 1.
Benford’s Law: Newcomb (1881), Benford (1938)

Statement

For many data sets, probability of observing a first digit of \(d\) base \(B\) is \(\log_B \left( \frac{d+1}{d} \right)\); base 10 about 30% are 1s.

- Not all data sets satisfy Benford’s Law.
  - Long street \([1, L]\): \(L = 199\) versus \(L = 999\).
  - Oscillates between 1/9 and 5/9 with first digit 1.
  - Many streets of different sizes: close to Benford.
Examples

- recurrence relations
- special functions (such as $n!$)
- iterates of power, exponential, rational maps
- products of random variables
- $L$-functions, characteristic polynomials
- iterates of the $3x + 1$ map
- differences of order statistics
- hydrology and financial data
- many hierarchical Bayesian models
Applications

- analyzing round-off errors
- determining the optimal way to store numbers
- detecting tax and image fraud, and data integrity
General Theory
Mantissas

Mantissa: $x = M_{10}(x) \cdot 10^k$, $k$ integer.
Mantissa: $x = M_{10}(x) \cdot 10^k$, $k$ integer.

$M_{10}(x) = M_{10}(\tilde{x})$ if and only if $x$ and $\tilde{x}$ have the same leading digits.
Mantissas

Mantissa: $x = M_{10}(x) \cdot 10^k$, $k$ integer.

$M_{10}(x) = M_{10}(\tilde{x})$ if and only if $x$ and $\tilde{x}$ have the same leading digits.

**Key observation:** $\log_{10}(x) = \log_{10}(\tilde{x}) \mod 1$ if and only if $x$ and $\tilde{x}$ have the same leading digits. Thus often study $y = \log_{10} x$. 
Equidistribution and Benford’s Law

**Equidistribution**

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

\[
\frac{\#\{n \leq N : y_n \mod 1 \in [a, b]\}}{N} \to b - a.
\]
Equidistribution

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

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

- Thm: \( \beta \not\in \mathbb{Q} \), \( n\beta \) is equidistributed mod 1.
Equidistribution and Benford’s Law

**Equidistribution**

\[ \{y_n\}_{n=1}^{\infty} \] is equidistributed modulo 1 if probability

\[ y_n \mod 1 \in [a, b] \] tends to \( b - a \):

\[
\frac{\#\{n \leq N : y_n \mod 1 \in [a, b]\}}{N} \to b - a.
\]

- Thm: \( \beta \notin \mathbb{Q}, n\beta \) is equidistributed mod 1.

- Examples: \( \log_{10} 2, \log_{10} \left(\frac{1+\sqrt{5}}{2}\right) \notin \mathbb{Q} \).
Equidistribution and Benford’s Law

Equidistribution

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

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

- Thm: \( \beta \notin \mathbb{Q} \), \( n\beta \) is equidistributed mod 1.

- Examples: \( \log_{10} 2 \), \( \log_{10} \left( \frac{1+\sqrt{5}}{2} \right) \notin \mathbb{Q} \).

**Proof:** if rational: \( 2 = 10^{p/q} \).
Equidistribution and Benford’s Law

Equidistribution

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

\[
\frac{\# \{ n \leq N : y_n \mod 1 \in [a, b]\}}{N} \to b - a.
\]

- Thm: \( \beta \notin \mathbb{Q}, n\beta \) is equidistributed mod 1.

- Examples: \( \log_{10} 2, \log_{10} \left( \frac{1+\sqrt{5}}{2} \right) \notin \mathbb{Q} \).

Proof: if rational: \( 2 = 10^{p/q} \).

Thus \( 2^q = 10^p \) or \( 2^{q-p} = 5^p \), impossible.
Example of Equidistribution: $n \sqrt{\pi} \mod 1$ for $n \leq 10$
Example of Equidistribution: $n\sqrt{\pi} \mod 1$

$n\sqrt{\pi} \mod 1$ for $n \leq 100$
Example of Equidistribution: $n\sqrt{\pi} \mod 1$

$n\sqrt{\pi} \mod 1$ for $n \leq 1000$
Example of Equidistribution: $n\sqrt{\pi} \mod 1$

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

Dense

A sequence \( \{z_n\}_{n=1}^{\infty} \) of numbers in \([0, 1]\) is dense if for any interval \([a, b]\) there are infinitely many \(z_n\) in \([a, b]\).

- Dirichlet’s Box (or Pigeonhole) Principle:
  If \(n + 1\) objects are placed in \(n\) boxes, at least one box has two objects.

- Denseness of \(n\alpha\):
  Thm: If \(\alpha \notin \mathbb{Q}\) then \(z_n = n\alpha \mod 1\) is dense.
**Proof** \( n\alpha \mod 1 \text{ dense if } \alpha \not\in \mathbb{Q} \)

- Enough to show in \([0, b]\) infinitely often for any \(b\).
- Choose any integer \(Q > 1/b\).
- \(Q\) bins: \([0, \frac{1}{Q}], [\frac{1}{Q}, \frac{2}{Q}], \ldots, [\frac{Q-1}{Q}, Q]\).
- \(Q + 1\) objects:
  \(\{\alpha \mod 1, 2\alpha \mod 1, \ldots, (Q+1)\alpha \mod 1\}\).
- Two in same bin, say \(q_1\alpha \mod 1\) and \(q_2\alpha \mod 1\).
- Exists integer \(p\) with \(0 < q_2\alpha - q_1\alpha - p < \frac{1}{Q}\).
- Get \((q_2 - q_1)\alpha \mod 1 \in [0, b]\).
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 \( y_i = \log_B x_i \).
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 \( y_i = \log_B x_i \).

Proof:

- \( x = M_B(x) \cdot B^k \) for some \( k \in \mathbb{Z} \).
- \( \text{FD}_B(x) = d \) iff \( d \leq M_B(x) < d + 1 \).
- \( \log_B d \leq y < \log_B(d + 1) \), \( y = \log_B x \) mod 1.
- If \( Y \sim \text{Unif}(0, 1) \) then above probability is \( \log_B \left( \frac{d+1}{d} \right) \).
Examples

- $2^n$ is Benford base 10 as $\log_{10} 2 \not\in \mathbb{Q}$. 
Examples

- Fibonacci numbers are Benford base 10.
Examples

- Fibonacci numbers are Benford base 10.

\[ a_{n+1} = a_n + a_{n-1}. \]
Examples

Fibonacci numbers are Benford base 10.

\[ a_{n+1} = a_n + a_{n-1}. \]

Guess \( a_n = n^r: r^{n+1} = r^n + r^{n-1} \) or \( r^2 = r + 1 \).
Examples

- Fibonacci numbers are Benford base 10.

\[ a_{n+1} = a_n + a_{n-1}. \]

Guess \( a_n = n^r: r^{n+1} = r^n + r^{n-1} \) or \( r^2 = r + 1 \).

Roots \( r = (1 \pm \sqrt{5})/2 \).
Examples

- Fibonacci numbers are Benford base 10.

$$a_{n+1} = a_n + a_{n-1}.$$  

Guess $$a_n = n^r: \ r^{n+1} = r^n + r^{n-1}$$ or $$r^2 = r + 1.$$  

Roots $$r = (1 \pm \sqrt{5})/2.$$  

General solution: $$a_n = c_1 r_1^n + c_2 r_2^n.$$
Fibonacci numbers are Benford base 10.

\[ a_{n+1} = a_n + a_{n-1}. \]

Guess \( a_n = n^r: \) \( r^{n+1} = r^n + r^{n-1} \) or \( r^2 = r + 1. \)

Roots \( r = (1 \pm \sqrt{5})/2. \)

General solution: \( a_n = c_1 r_1^n + c_2 r_2^n. \)

Binet: \( a_n = \frac{1}{\sqrt{5}} \left( \frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left( \frac{1-\sqrt{5}}{2} \right)^n. \)
Examples

- Fibonacci numbers are Benford base 10.
  \[ a_{n+1} = a_n + a_{n-1}. \]
  Guess \( a_n = n^r: \ r^{n+1} = r^n + r^{n-1} \) or \( r^2 = r + 1. \)
  Roots \( r = (1 \pm \sqrt{5})/2. \)
  General solution: \( a_n = c_1 r_1^n + c_2 r_2^n. \)
  Binet: \( a_n = \frac{1}{\sqrt{5}} \left( \frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left( \frac{1-\sqrt{5}}{2} \right)^n. \)

- Most linear recurrence relations Benford:
Examples

Fibonacci numbers are Benford base 10.

\[ a_{n+1} = a_n + a_{n-1}. \]

Guess \( a_n = n^r: r^{n+1} = r^n + r^{n-1} \) or \( r^2 = r + 1 \).

Roots \( r = (1 \pm \sqrt{5})/2 \).

General solution: \( a_n = c_1 r_1^n + c_2 r_2^n \).

Binet: \( a_n = \frac{1}{\sqrt{5}} \left( \frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left( \frac{1-\sqrt{5}}{2} \right)^n \).

Most linear recurrence relations Benford:

\[ a_{n+1} = 2a_n \]
Examples

- Fibonacci numbers are Benford base 10.
  \[ a_{n+1} = a_n + a_{n-1}. \]
  Guess \( a_n = n^r: \ r^{n+1} = r^n + r^{n-1} \) or \( r^2 = r + 1. \)
  Roots \( r = (1 \pm \sqrt{5})/2. \)
  General solution: \( a_n = c_1 r_1^n + c_2 r_2^n. \)
  Binet: \( a_n = \frac{1}{\sqrt{5}} \left( \frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left( \frac{1-\sqrt{5}}{2} \right)^n. \)

- Most linear recurrence relations Benford:
  \[ a_{n+1} = 2a_n - a_{n-1} \]
Examples

- Fibonacci numbers are Benford base 10.
  \[ a_{n+1} = a_n + a_{n-1}. \]
  Guess \( a_n = n^r: \ r^{n+1} = r^n + r^{n-1} \) or \( r^2 = r + 1 \).
  Roots \( r = (1 \pm \sqrt{5})/2 \).
  General solution: \( a_n = c_1 r_1^n + c_2 r_2^n \).
  Binet: \( a_n = \frac{1}{\sqrt{5}} \left( \frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left( \frac{1-\sqrt{5}}{2} \right)^n. \)

- Most linear recurrence relations Benford:
  - \( a_{n+1} = 2a_n - a_{n-1} \)
  - take \( a_0 = a_1 = 1 \) or \( a_0 = 0, \ a_1 = 1 \).
## Digits of $2^n$

First 60 values of $2^n$ (only displaying 30)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th>digit</th>
<th>#</th>
<th>Obs Prob</th>
<th>Benf Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1024</td>
<td>1048576</td>
<td>1</td>
<td>18</td>
<td>.300</td>
<td>.301</td>
</tr>
<tr>
<td>2</td>
<td>2048</td>
<td>2097152</td>
<td>2</td>
<td>12</td>
<td>.200</td>
<td>.176</td>
</tr>
<tr>
<td>4</td>
<td>4096</td>
<td>4194304</td>
<td>3</td>
<td>6</td>
<td>.100</td>
<td>.125</td>
</tr>
<tr>
<td>8</td>
<td>8192</td>
<td>8388608</td>
<td>4</td>
<td>6</td>
<td>.100</td>
<td>.097</td>
</tr>
<tr>
<td>16</td>
<td>16384</td>
<td>16777216</td>
<td>5</td>
<td>6</td>
<td>.100</td>
<td>.079</td>
</tr>
<tr>
<td>32</td>
<td>32768</td>
<td>33554432</td>
<td>6</td>
<td>4</td>
<td>.067</td>
<td>.067</td>
</tr>
<tr>
<td>64</td>
<td>65536</td>
<td>67108864</td>
<td>7</td>
<td>2</td>
<td>.033</td>
<td>.058</td>
</tr>
<tr>
<td>128</td>
<td>131072</td>
<td>134217728</td>
<td>8</td>
<td>5</td>
<td>.083</td>
<td>.051</td>
</tr>
<tr>
<td>256</td>
<td>262144</td>
<td>268435456</td>
<td>9</td>
<td>1</td>
<td>.017</td>
<td>.046</td>
</tr>
<tr>
<td>512</td>
<td>524288</td>
<td>536870912</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Analysis

- $\chi^2$-Tests: Test if theory describes data
Data Analysis

- **χ²-Tests**: Test if theory describes data
- **Expected probability**: $p_d = \log_{10} \left( \frac{d+1}{d} \right)$. 
Data Analysis

- $\chi^2$-Tests: Test if theory describes data
  - Expected probability: $p_d = \log_{10} \left( \frac{d+1}{d} \right)$.
  - Expect about $Np_d$ will have first digit $d$.
Data Analysis

- $\chi^2$-Tests: Test if theory describes data
  - Expected probability: $p_d = \log_{10}\left(\frac{d+1}{d}\right)$.
  - Expect about $Np_d$ will have first digit $d$.
  - Observe $\text{Obs}(d)$ with first digit $d$. 
Data Analysis

\( \chi^2 \)-Tests: Test if theory describes data

- Expected probability: \( p_d = \log_{10} \left( \frac{d+1}{d} \right) \).
- Expect about \( Np_d \) will have first digit \( d \).
- Observe \( \text{Obs}(d) \) with first digit \( d \).

\[
\chi^2 = \sum_{d=1}^{9} \frac{(\text{Obs}(d) - Np_d)^2}{Np_d}.
\]
Data Analysis

- $\chi^2$-Tests: Test if theory describes data
  - Expected probability: $p_d = \log_{10} \left( \frac{d+1}{d} \right)$.
  - Expect about $Np_d$ will have first digit $d$.
  - Observe $\text{Obs}(d)$ with first digit $d$.
  - $\chi^2 = \sum_{d=1}^{9} \frac{(\text{Obs}(d) - Np_d)^2}{Np_d}$.
  - Smaller $\chi^2$, more likely correct model.
Data Analysis

\( \chi^2 \)-Tests: Test if theory describes data
- Expected probability: \( p_d = \log_{10} \left( \frac{d+1}{d} \right) \).
- Expect about \( Np_d \) will have first digit \( d \).
- Observe \( \text{Obs}(d) \) with first digit \( d \).
- \( \chi^2 = \sum_{d=1}^{9} \frac{(\text{Obs}(d) - Np_d)^2}{Np_d} \).
- Smaller \( \chi^2 \), more likely correct model.

Will study \( \gamma^n, e^n, \pi^n \).
Logarithms and Benford’s Law

\( \chi^2 \) values for \( \alpha^n, 1 \leq n \leq N \) (5% 15.5).

<table>
<thead>
<tr>
<th>( N )</th>
<th>( \chi^2(\gamma) )</th>
<th>( \chi^2(e) )</th>
<th>( \chi^2(\pi) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.72</td>
<td>0.30</td>
<td>46.65</td>
</tr>
<tr>
<td>200</td>
<td>0.24</td>
<td>0.30</td>
<td>8.58</td>
</tr>
<tr>
<td>400</td>
<td>0.14</td>
<td>0.10</td>
<td>10.55</td>
</tr>
<tr>
<td>500</td>
<td>0.08</td>
<td>0.07</td>
<td>2.69</td>
</tr>
<tr>
<td>700</td>
<td>0.19</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>800</td>
<td>0.04</td>
<td>0.03</td>
<td>6.19</td>
</tr>
<tr>
<td>900</td>
<td>0.09</td>
<td>0.09</td>
<td>1.71</td>
</tr>
<tr>
<td>1000</td>
<td>0.02</td>
<td>0.06</td>
<td>2.90</td>
</tr>
</tbody>
</table>
Logarithms and Benford’s Law: Base 10

\[ \log(\chi^2) \text{ vs } N \text{ for } \pi^n \text{ (red)} \text{ and } e^n \text{ (blue)}, \quad n \in \{1, \ldots, N\}. \text{ Note } \pi^{175} \approx 1.0028 \cdot 10^{87}, (5\%, \log(\chi^2) \approx 2.74). \]
Logarithms and Benford’s Law: Base 20

\[ \log(\chi^2) \text{ vs } N \text{ for } \pi^n \text{ (red) and } e^n \text{ (blue)}, \]
\[ n \in \{1, \ldots, N\}. \text{ Note } e^3 \approx 20.0855, \text{ (5\%, } \log(\chi^2) \approx 2.74). \]
Applications
## Stock Market

<table>
<thead>
<tr>
<th>Milestone</th>
<th>Date</th>
<th>Effective Rate from last milestone</th>
</tr>
</thead>
<tbody>
<tr>
<td>108.35</td>
<td>Jan 12, 1906</td>
<td>3.0%</td>
</tr>
<tr>
<td>500.24</td>
<td>Mar 12, 1956</td>
<td>4.2%</td>
</tr>
<tr>
<td>1003.16</td>
<td>Nov 14, 1972</td>
<td>4.9%</td>
</tr>
<tr>
<td>2002.25</td>
<td>Jan 8, 1987</td>
<td>9.5%</td>
</tr>
<tr>
<td>3004.46</td>
<td>Apr 17, 1991</td>
<td>7.4%</td>
</tr>
<tr>
<td>4003.33</td>
<td>Feb 23, 1995</td>
<td>30.6%</td>
</tr>
<tr>
<td>5023.55</td>
<td>Nov 21, 1995</td>
<td>20.0%</td>
</tr>
<tr>
<td>6010.00</td>
<td>Oct 14, 1996</td>
<td>46.6%</td>
</tr>
<tr>
<td>7022.44</td>
<td>Feb 13, 1997</td>
<td>32.3%</td>
</tr>
<tr>
<td>8038.88</td>
<td>Jul 16, 1997</td>
<td>16.1%</td>
</tr>
<tr>
<td>9033.23</td>
<td>Apr 6, 1998</td>
<td>10.5%</td>
</tr>
<tr>
<td>10006.78</td>
<td>Mar 29, 1999</td>
<td>38.0%</td>
</tr>
<tr>
<td>11209.84</td>
<td>Jul 16, 1999</td>
<td>1.0%</td>
</tr>
<tr>
<td>12011.73</td>
<td>Oct 19, 2006</td>
<td>16.7%</td>
</tr>
<tr>
<td>13089.89</td>
<td>Apr 25, 2007</td>
<td>28.9%</td>
</tr>
<tr>
<td>14000.41</td>
<td>Jul 19, 2007</td>
<td></td>
</tr>
</tbody>
</table>
Applications for the IRS: Detecting Fraud
Applications for the IRS: Detecting Fraud
Applications for the IRS: Detecting Fraud

Exhibit 3: Check Fraud in Arizona

The table lists the checks that a manager in the office of the Arizona State Treasurer wrote to divert funds for his own use. The vendors to whom the checks were issued were fictitious.

<table>
<thead>
<tr>
<th>Date of Check</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>October 9, 1992</td>
<td>$ 1,927.48</td>
</tr>
<tr>
<td></td>
<td>27,902.31</td>
</tr>
<tr>
<td>October 14, 1992</td>
<td>86,241.90</td>
</tr>
<tr>
<td></td>
<td>72,117.46</td>
</tr>
<tr>
<td></td>
<td>81,321.75</td>
</tr>
<tr>
<td></td>
<td>97,473.96</td>
</tr>
<tr>
<td>October 19, 1992</td>
<td>93,249.11</td>
</tr>
<tr>
<td></td>
<td>89,658.17</td>
</tr>
<tr>
<td></td>
<td>87,776.89</td>
</tr>
<tr>
<td></td>
<td>92,105.83</td>
</tr>
<tr>
<td></td>
<td>79,949.16</td>
</tr>
<tr>
<td></td>
<td>87,602.93</td>
</tr>
<tr>
<td></td>
<td>96,879.27</td>
</tr>
<tr>
<td></td>
<td>91,806.47</td>
</tr>
<tr>
<td></td>
<td>84,991.67</td>
</tr>
<tr>
<td></td>
<td>90,831.83</td>
</tr>
<tr>
<td></td>
<td>93,766.67</td>
</tr>
<tr>
<td></td>
<td>88,338.72</td>
</tr>
<tr>
<td></td>
<td>94,639.49</td>
</tr>
<tr>
<td></td>
<td>83,709.28</td>
</tr>
<tr>
<td></td>
<td>96,412.21</td>
</tr>
<tr>
<td></td>
<td>88,432.86</td>
</tr>
<tr>
<td></td>
<td>71,552.16</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>$ 1,878,687.58</strong></td>
</tr>
</tbody>
</table>
Applications for the IRS: Detecting Fraud (cont)

- Embezzler started small and then increased dollar amounts.

- Most amounts below $100,000 (critical threshold for data requiring additional scrutiny).

- Over 90% had first digit of 7, 8 or 9.
Detecting Fraud

Audit of a bank revealed huge spike of numbers starting with 48 and 49, most due to one person.

Write-off limit of $5,000. Officer had friends applying for credit cards, ran up balances just under $5,000 then he would write the debts off.
Detecting Fraud

Enron

- Benford’s Law detected manipulation of revenue numbers.

- Results showed a tendency towards round Earnings Per Share (0.10, 0.20, etc.). Consistent with a small but noticeable increase in earnings management in 2002.
Data Integrity: Stream Flow Statistics: 130 years, 457,440 records
Benford Good Processes
Poisson Summation and Benford’s Law: Definitions

- Feller, Pinkham (often exact processes)
Poisson Summation and Benford’s Law: Definitions

- Feller, Pinkham (often exact processes)
- data $Y_{T,B} = \log_B \overrightarrow{X}_T$ (discrete/continuous):

$$\mathbb{P}(A) = \lim_{T \to \infty} \frac{\#\{n \in A : n \leq T\}}{T}$$
Poisson Summation and Benford’s Law: Definitions

- Feller, Pinkham (often exact processes)
- data \( Y_{T,B} = \log_B \vec{X}_T \) (discrete/continuous):

\[
\mathbb{P}(A) = \lim_{T \to \infty} \frac{\#\{n \in A : n \leq T\}}{T}
\]

- Poisson Summation Formula: \( f \) nice:

\[
\sum_{\ell=-\infty}^{\infty} f(\ell) = \sum_{\ell=-\infty}^{\infty} \hat{f}(\ell),
\]

Fourier transform \( \hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)e^{-2\pi i x \xi} \, dx \).
Benford Good Process

\( X_T \) is Benford Good if there is a nice \( f \) st

\[
\text{CDF}_{Y_{T,B}}(y) = \int_{-\infty}^{y} \frac{1}{T} f \left( \frac{t}{T} \right) \, dt + E_T(y) := G_T(y)
\]

and monotonically increasing \( h (h(|T|) \to \infty) :\)
**Benford Good Process**

$X_T$ is **Benford Good** if there is a nice $f$ st

$$\text{CDF}_{Y_{T,B}}(y) = \int_{-\infty}^{y} \frac{1}{T} f \left( \frac{t}{T} \right) dt + E_T(y) := G_T(y)$$

and monotonically increasing $h (h(|T|) \to \infty)$:

- **Small tails:** $G_T(\infty) - G_T(Th(T)) = o(1)$,
  $G_T(-Th(T)) - G_T(-\infty) = 0(1)$. 


Benford Good Process

$X_T$ is Benford Good if there is a nice CDF $F_{Y_T,B}(y) = \int_{-\infty}^{y} \frac{1}{T} f \left( \frac{t}{T} \right) dt + E_T(y) := G_T(y)$

and monotonically increasing $h (h(|T|) \to \infty)$:

- **Small tails:** $G_T(\infty) - G_T(Th(T)) = o(1)$, $G_T(-Th(T)) - G_T(-\infty) = 0(1)$.

- **Decay of the Fourier Transform:** $\sum_{\ell \neq 0} \left| \frac{\hat{f}(T\ell)}{\ell} \right| = o(1)$. 
Benford Good Process

$X_T$ is Benford Good if there is a nice CDF

$$\text{CDF}_{Y_{T,B}}(y) = \int_{-\infty}^{y} \frac{1}{T} f \left( \frac{t}{T} \right) dt + E_T(y) \equiv G_T(y)$$

and monotonically increasing $h (h(|T|) \to \infty)$:

- **Small tails:** $G_T(\infty) - G_T(Th(T)) = o(1)$, $G_T(-Th(T)) - G_T(-\infty) = 0(1)$.

- **Decay of the Fourier Transform:** $\sum_{\ell \neq 0} \left| \frac{\hat{f}(T\ell)}{\ell} \right| = o(1)$.

- **Small translated error:** $\mathcal{E}(a, b, T)) = \sum_{|\ell| \leq Th(T)} [E_T(b + \ell) - E_T(a + \ell)] = o(1)$. 

Main Theorem

**Theorem (Kontorovich and M–, 2005)**

\[ X_T \text{ converging to } X \text{ as } T \to \infty \text{ (think spreading Gaussian)}. \text{ If } X_T \text{ is Benford good, then } X \text{ is Benford.} \]
Main Theorem

**Theorem (Kontorovich and M–, 2005)**

\[ X_T \text{ converging to } X \text{ as } T \to \infty \text{ (think spreading Gaussian). If } X_T \text{ is Benford good, then } X \text{ is Benford.} \]

- **Examples**
  - *L*-functions
  - characteristic polynomials (RMT)
  - $3x + 1$ problem
  - geometric Brownian motion.
Sketch of the proof

Structure Theorem:
- main term is something nice spreading out
- apply Poisson summation
Sketch of the proof

Structure Theorem:
- main term is something nice spreading out
- apply Poisson summation

Control translated errors:
- hardest step
- techniques problem specific
Sketch of the proof (continued)

\[
\sum_{\ell = -\infty}^{\infty} \mathbb{P} \left( a + \ell \leq \bar{Y}_{T,B} \leq b + \ell \right)
\]
Sketch of the proof (continued)

\[
\sum_{\ell=-\infty}^{\infty} \Pr \left( a + \ell \leq \bar{Y}_{T,B} \leq b + \ell \right) = \sum_{|\ell| \leq Th(T)} [G_T(b + \ell) - G_T(a + \ell)] + o(1)
\]
Sketch of the proof (continued)

\[
\sum_{\ell=-\infty}^{\infty} \mathbb{P}\left(a + \ell \leq \tilde{Y}_{T,B} \leq b + \ell\right)
\]

\[
= \sum_{|\ell| \leq Th(T)} [G_T(b + \ell) - G_T(a + \ell)] + o(1)
\]

\[
= \int_{a}^{b} \sum_{|\ell| \leq Th(T)} \frac{1}{T} \cdot f\left(\frac{t}{T}\right) dt + \mathcal{E}(a, b, T) + o(1)
\]
Sketch of the proof (continued)

\[
\sum_{\ell = -\infty}^{\infty} \mathbb{P}\left(a + \ell \leq \hat{Y}_{T,B} \leq b + \ell\right) = \sum_{|\ell| \leq Th(T)} [G_T(b + \ell) - G_T(a + \ell)] + o(1)
\]

\[
= \int_{a}^{b} \sum_{|\ell| \leq Th(T)} \frac{1}{T} f\left(\frac{t}{T}\right) dt + \mathcal{E}(a, b, T) + o(1)
\]

\[
= \hat{f}(0) \cdot (b - a) + \sum_{\ell \neq 0} \hat{f}(T\ell) \frac{e^{2\pi i b\ell} - e^{2\pi i a\ell}}{2\pi i \ell} + o(1).
\]
Riemann Zeta Function

\[ \zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{p \text{ prime}} \left(1 - \frac{1}{p^s}\right)^{-1}. \]
Riemann Zeta Function

\[ \zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{p \text{ prime}} \left( 1 - \frac{1}{p^s} \right)^{-1}. \]

\[ \prod_{p \text{ prime}} \left( 1 - \frac{1}{p^s} \right)^{-1} = \prod_{p \text{ prime}} \left( 1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \cdots \right) \]

\[ = \left( 1 + \frac{1}{2^s} + \frac{1}{2^{2s}} + \cdots \right) \left( 1 + \frac{1}{3^s} + \frac{1}{3^{2s}} + \cdots \right) \]

\[ = 1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \frac{1}{5^s} + \frac{1}{(2 \cdot 3)^s} + \cdots \]
Riemann Zeta Function

\[ \zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{\text{prime } p} \left( 1 - \frac{1}{p^s} \right)^{-1}. \]

\[ \lim_{s \to 1^+} \zeta(s) = \infty \text{ implies infinitely many primes.} \]
Riemann Zeta Function

\[ \zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{\text{prime } p} \left(1 - \frac{1}{p^s}\right)^{-1}. \]

\[ \lim_{s \to 1^+} \zeta(s) = \infty \text{ implies infinitely many primes.} \]

\[ \zeta(2) = \frac{\pi^2}{6} \text{ implies infinitely many primes.} \]
Riemann Zeta Function

\[ \left| \zeta \left( \frac{1}{2} + i \frac{k}{4} \right) \right| , \ k \in \{0, 1, \ldots, 65535\}. \]
The 3x + 1 Problem and Benford’s Law
Kakutani (conspiracy), Erdös (not ready).

x odd, $T(x) = \frac{3x + 1}{2^k}$, $2^k \parallel 3x + 1$. 
3x + 1 Problem

- Kakutani (conspiracy), Erdös (not ready).

- x odd, $T(x) = \frac{3x+1}{2^k}$, $2^k || 3x + 1$.

- Conjecture: for some $n = n(x)$, $T^n(x) = 1$. 
3x + 1 Problem

- Kakutani (conspiracy), Erdős (not ready).
- x odd, $T(x) = \frac{3x+1}{2^k}, 2^k || 3x + 1$.
- Conjecture: for some $n = n(x)$, $T^n(x) = 1$.
- 7
Kakutani (conspiracy), Erdös (not ready).

- $x$ odd, $T(x) = \frac{3x+1}{2^k}$, $2^k || 3x + 1$.

- Conjecture: for some $n = n(x)$, $T^n(x) = 1$.

- $7 \rightarrow 11$
3x + 1 Problem

- Kakutani (conspiracy), Erdös (not ready).

- $x$ odd, \( T(x) = \frac{3x+1}{2^k}, \ 2^k || 3x + 1. \)

- Conjecture: for some \( n = n(x), \ T^n(x) = 1. \)

- 7 \( \rightarrow_1 11 \rightarrow_1 17 \)
3x + 1 Problem

- Kakutani (conspiracy), Erdös (not ready).

- x odd, \( T(x) = \frac{3x+1}{2^k}, \ 2^k || 3x + 1. \)

- Conjecture: for some \( n = n(x), \ T^n(x) = 1. \)

- 7 \( \rightarrow_1 11 \rightarrow_1 17 \rightarrow_2 13 \)
Kakutani (conspiracy), Erdös (not ready).

$x$ odd, $T(x) = \frac{3x+1}{2^k}$, $2^k \mid |3x + 1|$.

Conjecture: for some $n = n(x)$, $T^n(x) = 1$.

$7 \rightarrow_1 11 \rightarrow_1 17 \rightarrow_2 13 \rightarrow_3 5$
Kakutani (conspiracy), Erdös (not ready).

$x$ odd, $T(x) = \frac{3x+1}{2^k}$, $2^k || 3x + 1$.

Conjecture: for some $n = n(x)$, $T^n(x) = 1$.

$7 \rightarrow_1 11 \rightarrow_1 17 \rightarrow_2 13 \rightarrow_3 5 \rightarrow_4 1$
3x + 1 Problem

- Kakutani (conspiracy), Erdös (not ready).
- x odd, \( T(x) = \frac{3x+1}{2^k}, 2^k \mid |3x + 1| \).
- Conjecture: for some \( n = n(x), T^n(x) = 1 \).
- 7 →_1 11 →_1 17 →_2 13 →_3 5 →_4 1 →_2 1,
3x + 1 Problem

- Kakutani (conspiracy), Erdös (not ready).

- x odd, \( T(x) = \frac{3x+1}{2^k}, 2^k \| 3x + 1 \).

- Conjecture: for some \( n = n(x), T^n(x) = 1 \).

- 7 \( \rightarrow_1 \) 11 \( \rightarrow_1 \) 17 \( \rightarrow_2 \) 13 \( \rightarrow_3 \) 5 \( \rightarrow_4 \) 1 \( \rightarrow_2 \) 1,
  2-path (1, 1), 5-path (1, 1, 2, 3, 4).
  \( m \)-path: \( (k_1, \ldots, k_m) \).
Heuristic Proof of $3x + 1$ Conjecture

$$a_{n+1} = T(a_n)$$
Heuristic Proof of $3x + 1$ Conjecture

\[ a_{n+1} = T(a_n) \]

\[ \mathbb{E}[\log a_{n+1}] \approx \sum_{k=1}^{\infty} \frac{1}{2^k} \log \left( \frac{3a_n}{2^k} \right) \]
Heuristic Proof of $3x + 1$ Conjecture

\[
a_{n+1} = T(a_n)
\]

\[
\mathbb{E}[\log a_{n+1}] \approx \sum_{k=1}^{\infty} \frac{1}{2^k} \log \left( \frac{3a_n}{2^k} \right)
\]

\[
= \log a_n + \log 3 - \log 2 \sum_{k=1}^{\infty} \frac{k}{2^k}
\]
Heuristic Proof of $3x + 1$ Conjecture

\[ a_{n+1} = T(a_n) \]

\[ \mathbb{E}[\log a_{n+1}] \approx \sum_{k=1}^{\infty} \frac{1}{2^k} \log \left( \frac{3a_n}{2^k} \right) \]

\[ = \log a_n + \log 3 - \log 2 \sum_{k=1}^{\infty} \frac{k}{2^k} \]

\[ = \log a_n + \log \left( \frac{3}{4} \right) . \]

Geometric Brownian Motion, drift $\log(3/4) < 1$. 
Structure Theorem: Sinai, Kontorovich-Sinai

\[ P(A) = \lim_{N \to \infty} \frac{\#\{n \leq N: n \equiv 1, 5 \mod 6, n \in A\}}{\#\{n \leq N: n \equiv 1, 5 \mod 6\}}. \]
Structure Theorem: Sinai, Kontorovich-Sinai

\[ P(A) = \lim_{N \to \infty} \frac{\# \{ n \leq N : n \equiv 1, 5 \mod 6, n \in A \}}{\# \{ n \leq N : n \equiv 1, 5 \mod 6 \}}. \]

\((k_1, \ldots, k_m)\): two full arithm progressions:

\[ 6 \cdot 2^{k_1 + \cdots + k_m} p + q. \]
Structure Theorem: Sinai, Kontorovich-Sinai

\[ \mathbb{P}(A) = \lim_{N \to \infty} \frac{\# \{ n \leq N : n \equiv 1, 5 \text{ mod } 6, n \in A \}}{\# \{ n \leq N : n \equiv 1, 5 \text{ mod } 6 \}}. \]

\((k_1, \ldots, k_m)\): two full arithm progressions:
\[ 6 \cdot 2^{k_1 + \cdots + k_m} p + q. \]

**Theorem (Sinai, Kontorovich-Sinai)**

\(k_i\)-values are i.i.d.r.v. \((\text{geometric, } 1/2)\):

\[ \mathbb{P} \left( \log_2 \left[ \frac{x_m}{\left( \frac{3}{4} \right)^m x_0} \right] \leq a \right) = \mathbb{P} \left( \frac{S_m - 2m}{\sqrt{2m}} \leq a \right) \]
Structure Theorem: Sinai, Kontorovich-Sinai

\[ P(A) = \lim_{N \to \infty} \frac{\#\{n \leq N : n \equiv 1, 5 \mod 6, n \in A\}}{\#\{n \leq N : n \equiv 1, 5 \mod 6\}} . \]

\((k_1, \ldots, k_m)\): two full arithm progressions:
\[ 6 \cdot 2^{k_1} + \cdots + 2^{k_m} p + q. \]

Theorem (Sinai, Kontorovich-Sinai)

\(k_i\)-values are i.i.d.r.v. (geometric, 1/2):

\[
P \left( \log_2 \left[ \frac{x_m}{\left(\frac{3}{4}\right)^m x_0} \right] \leq a \right) = P \left( \frac{S_m - 2m}{(\log_2 B) \sqrt{2m}} \leq a \right)
\]
Structure Theorem: Sinai, Kontorovich-Sinai

\[ P(A) = \lim_{N \to \infty} \frac{\# \{ n \leq N : n \equiv 1,5 \text{ mod } 6, n \in A \}}{\# \{ n \leq N : n \equiv 1,5 \text{ mod } 6 \}} . \]

\((k_1, \ldots, k_m)\): two full arithm progressions:

\[ 6 \cdot 2^{k_1 + \cdots + k_m} p + q. \]

### Theorem (Sinai, Kontorovich-Sinai)

\(k_i\)-values are i.i.d.r.v. (geometric, 1/2):

\[ P \left( \frac{\log_B \left[ \frac{x_m}{(\frac{3}{4})^m x_0} \right]}{\sqrt{2m}} \leq a \right) = P \left( \frac{(S_m - 2m)}{\log_2 B} \leq a \right) \]
3x + 1 and Benford

**Theorem (Kontorovitch and M–, 2005)**

As $m \to \infty$, $x_m/(3/4)^m x_0$ is Benford.

**Theorem (Lagarias-Soundararajan 2006)**

$X \geq 2^N$, for all but at most $c(B)N^{-1/36} X$ initial seeds the distribution of the first $N$ iterates of the $3x + 1$ map are within $2N^{-1/36}$ of the Benford probabilities.
Sketch of the proof

- Failed Proof: lattices, bad errors.
Sketch of the proof

- Failed Proof: lattices, bad errors.

- CLT: \( (S_m - 2m)/\sqrt{2m} \to N(0, 1) \):

  \[
  \mathbb{P} \left( S_m - 2m = k \right) = \frac{\eta(k/\sqrt{m})}{\sqrt{m}} + O \left( \frac{1}{g(m)\sqrt{m}} \right).
  \]
Sketch of the proof

- Failed Proof: lattices, bad errors.

- CLT: \((S_m - 2m)/\sqrt{2m} \to N(0, 1)\):

\[ P(S_m - 2m = k) = \frac{\eta(k/\sqrt{m})}{\sqrt{m}} + O\left(\frac{1}{g(m)\sqrt{m}}\right). \]

- Quantified Equidistribution:

\[ I_\ell = \{\ell M, \ldots, (\ell + 1)M - 1\}, \quad M = m^c, \quad c < 1/2 \]
Sketch of the proof

- Failed Proof: lattices, bad errors.

- CLT: \((S_m - 2m)/\sqrt{2m} \rightarrow N(0, 1)\):

\[ P(S_m - 2m = k) = \frac{\eta(k/\sqrt{m})}{\sqrt{m}} + O\left(\frac{1}{g(m)\sqrt{m}}\right). \]

- Quantified Equidistribution:

  \( I_\ell = \{\ell M, \ldots, (\ell + 1) M - 1\}, \quad M = m^c, \quad c < 1/2 \)

  \( k_1, k_2 \in I_\ell: \quad \left| \eta\left(\frac{k_1}{\sqrt{m}}\right) - \eta\left(\frac{k_2}{\sqrt{m}}\right) \right| \text{ small} \)
Sketch of the proof

- Failed Proof: lattices, bad errors.

- CLT: \((S_m - 2m)/\sqrt{2m} \to N(0, 1)\):

\[
P(S_m - 2m = k) = \frac{\eta(k/\sqrt{m})}{\sqrt{m}} + O\left(\frac{1}{g(m)\sqrt{m}}\right).
\]

- Quantified Equidistribution:

\[I_\ell = \{\ell M, \ldots, (\ell + 1)M - 1\}, \quad M = m^c, \quad c < 1/2\]

\[k_1, k_2 \in I_\ell: \quad \left|\eta\left(\frac{k_1}{\sqrt{m}}\right) - \eta\left(\frac{k_2}{\sqrt{m}}\right)\right| \text{ small}\]

\[C = \log_B 2 \text{ of irrationality type } \kappa < \infty:\]

\[
\#\{k \in I_\ell : \overline{kc} \in [a, b]\} = M(b - a) + O(M^{1+\epsilon-1/\kappa}).
\]
Irrationality Type

\( \alpha \) has irrationality type \( \kappa \) if \( \kappa \) is the supremum of all \( \gamma \) with

\[
\lim_{q \to \infty} q^{\gamma+1} \min_p \left| \alpha - \frac{p}{q} \right| = 0.
\]

- Algebraic irrationals: type 1 (Roth’s Thm).
- Theory of Linear Forms: \( \log_B 2 \) of finite type.
Linear Forms

**Theorem (Baker)**

\( \alpha_1, \ldots, \alpha_n \) algebraic numbers height \( A_j \geq 4, \beta_1, \ldots, \beta_n \in \mathbb{Q} \) with height at most \( B \geq 4 \),

\[ \Lambda = \beta_1 \log \alpha_1 + \cdots + \beta_n \log \alpha_n. \]

*If \( \Lambda \neq 0 \) then \( |\Lambda| > B^{-C \Omega \log \Omega'} \), with \( d = [\mathbb{Q}(\alpha_i, \beta_j) : \mathbb{Q}] \), \( C = (16nd)^{200n} \), \( \Omega = \prod_j \log A_j \), \( \Omega' = \Omega / \log A_n \).*

Gives \( \log_{10} 2 \) of finite type, with \( \kappa < 1.2 \cdot 10^{602} \):

\[ |\log_{10} 2 - p/q| = |q \log 2 - p \log 10| / q \log 10. \]
Theorem (Erdös-Turan)

\[ D_N = \sup_{[a,b]} \frac{|N(b-a) - \#\{n \leq N : x_n \in [a, b]\}|}{N} \]

There is a \( C \) such that for all \( m \):

\[ D_N \leq C \cdot \left( \frac{1}{m} + \sum_{h=1}^{m} \frac{1}{h} \left| \frac{1}{N} \sum_{n=1}^{N} e^{2\pi ihx_n} \right| \right) \]
Proof of Erdös-Turan

Consider special case $x_n = n\alpha$, $\alpha \notin \mathbb{Q}$.

- Exponential sum $\leq \frac{1}{|\sin(\pi h\alpha)|} \leq \frac{1}{2||h\alpha||}$.
- Must control $\sum_{h=1}^{m} \frac{1}{h||h\alpha||}$, see irrationality type enter.
- type $\kappa$, $\sum_{h=1}^{m} \frac{1}{h||h\alpha||} = O\left(m^{\kappa-1+\epsilon}\right)$, take $m = \lceil N^{1/\kappa} \rceil$. 
3x + 1 Data: random 10,000 digit number, $2^k \parallel 3x + 1$

80,514 iterations ($(4/3)^n = a_0$ predicts 80,319); $\chi^2 = 13.5$ (5% 15.5).

<table>
<thead>
<tr>
<th>Digit</th>
<th>Number</th>
<th>Observed</th>
<th>Benford</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24251</td>
<td>0.301</td>
<td>0.301</td>
</tr>
<tr>
<td>2</td>
<td>14156</td>
<td>0.176</td>
<td>0.176</td>
</tr>
<tr>
<td>3</td>
<td>10227</td>
<td>0.127</td>
<td>0.125</td>
</tr>
<tr>
<td>4</td>
<td>7931</td>
<td>0.099</td>
<td>0.097</td>
</tr>
<tr>
<td>5</td>
<td>6359</td>
<td>0.079</td>
<td>0.079</td>
</tr>
<tr>
<td>6</td>
<td>5372</td>
<td>0.067</td>
<td>0.067</td>
</tr>
<tr>
<td>7</td>
<td>4476</td>
<td>0.056</td>
<td>0.058</td>
</tr>
<tr>
<td>8</td>
<td>4092</td>
<td>0.051</td>
<td>0.051</td>
</tr>
<tr>
<td>9</td>
<td>3650</td>
<td>0.045</td>
<td>0.046</td>
</tr>
</tbody>
</table>
3x + 1 Data: random 10,000 digit number, 2|3x + 1

241,344 iterations, $\chi^2 = 11.4$ (5% 15.5).

<table>
<thead>
<tr>
<th>Digit</th>
<th>Number</th>
<th>Observed</th>
<th>Benford</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>72924</td>
<td>0.302</td>
<td>0.301</td>
</tr>
<tr>
<td>2</td>
<td>42357</td>
<td>0.176</td>
<td>0.176</td>
</tr>
<tr>
<td>3</td>
<td>30201</td>
<td>0.125</td>
<td>0.125</td>
</tr>
<tr>
<td>4</td>
<td>23507</td>
<td>0.097</td>
<td>0.097</td>
</tr>
<tr>
<td>5</td>
<td>18928</td>
<td>0.078</td>
<td>0.079</td>
</tr>
<tr>
<td>6</td>
<td>16296</td>
<td>0.068</td>
<td>0.067</td>
</tr>
<tr>
<td>7</td>
<td>13702</td>
<td>0.057</td>
<td>0.058</td>
</tr>
<tr>
<td>8</td>
<td>12356</td>
<td>0.051</td>
<td>0.051</td>
</tr>
<tr>
<td>9</td>
<td>11073</td>
<td>0.046</td>
<td>0.046</td>
</tr>
</tbody>
</table>
5x + 1 Data: random 10,000 digit number, \( 2^k \parallel 5x + 1 \)

27,004 iterations, \( \chi^2 = 1.8 \) (5% 15.5).

<table>
<thead>
<tr>
<th>Digit</th>
<th>Number</th>
<th>Observed</th>
<th>Benford</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8154</td>
<td>0.302</td>
<td>0.301</td>
</tr>
<tr>
<td>2</td>
<td>4770</td>
<td>0.177</td>
<td>0.176</td>
</tr>
<tr>
<td>3</td>
<td>3405</td>
<td>0.126</td>
<td>0.125</td>
</tr>
<tr>
<td>4</td>
<td>2634</td>
<td>0.098</td>
<td>0.097</td>
</tr>
<tr>
<td>5</td>
<td>2105</td>
<td>0.078</td>
<td>0.079</td>
</tr>
<tr>
<td>6</td>
<td>1787</td>
<td>0.066</td>
<td>0.067</td>
</tr>
<tr>
<td>7</td>
<td>1568</td>
<td>0.058</td>
<td>0.058</td>
</tr>
<tr>
<td>8</td>
<td>1357</td>
<td>0.050</td>
<td>0.051</td>
</tr>
<tr>
<td>9</td>
<td>1224</td>
<td>0.045</td>
<td>0.046</td>
</tr>
</tbody>
</table>
5x + 1 Data: random 10,000 digit number, 2 \mid 5x + 1

241,344 iterations, \( \chi^2 = 3 \cdot 10^{-4} \) (5% 15.5).

<table>
<thead>
<tr>
<th>Digit</th>
<th>Number</th>
<th>Observed</th>
<th>Benford</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>72652</td>
<td>0.301</td>
<td>0.301</td>
</tr>
<tr>
<td>2</td>
<td>42499</td>
<td>0.176</td>
<td>0.176</td>
</tr>
<tr>
<td>3</td>
<td>30153</td>
<td>0.125</td>
<td>0.125</td>
</tr>
<tr>
<td>4</td>
<td>23388</td>
<td>0.097</td>
<td>0.097</td>
</tr>
<tr>
<td>5</td>
<td>19110</td>
<td>0.079</td>
<td>0.079</td>
</tr>
<tr>
<td>6</td>
<td>16159</td>
<td>0.067</td>
<td>0.067</td>
</tr>
<tr>
<td>7</td>
<td>13995</td>
<td>0.058</td>
<td>0.058</td>
</tr>
<tr>
<td>8</td>
<td>12345</td>
<td>0.051</td>
<td>0.051</td>
</tr>
<tr>
<td>9</td>
<td>11043</td>
<td>0.046</td>
<td>0.046</td>
</tr>
</tbody>
</table>
Products and Chains of Random Variables
Key Ingredients

- Mellin transform and Fourier transform related by logarithmic change of variable.

- Poisson summation from collapsing to modulo 1 random variables.
Preliminaries

- $\Xi_1, \ldots, \Xi_n$ nice independent r.v.'s on $[0, \infty)$.
Preliminaries

- $\Xi_1, \ldots, \Xi_n$ nice independent r.v.'s on $[0, \infty)$.
- Density $\Xi_1 \cdot \Xi_2$:

$$\int_0^\infty f_2 \left( \frac{x}{t} \right) f_1(t) \frac{dt}{t}$$
Preliminaries

- $\xi_1, \ldots, \xi_n$ nice independent r.v.'s on $[0, \infty)$.
- Density $\xi_1 \cdot \xi_2$:

$$\int_0^\infty f_2 \left( \frac{x}{t} \right) f_1(t) \frac{dt}{t}$$

- Proof: $\text{Prob}(\xi_1 \cdot \xi_2 \in [0, x])$:

$$\int_{t=0}^\infty \text{Prob} \left( \xi_2 \in \left[ 0, \frac{x}{t} \right] \right) f_1(t) dt$$

$$= \int_{t=0}^\infty F_2 \left( \frac{x}{t} \right) f_1(t) dt,$$

differentiate.
Mellin Transform

\[(\mathcal{M} f)(s) = \int_{0}^{\infty} f(x) x^{s} \frac{dx}{x}\]

\[(\mathcal{M}^{-1} g)(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} g(s) x^{-s} ds\]

\[g(s) = (\mathcal{M} f)(s), \quad f(x) = (\mathcal{M}^{-1} g)(x).\]

\[(f_1 \ast f_2)(x) = \int_{0}^{\infty} f_2 \left( \frac{x}{t} \right) f_1(t) \frac{dt}{t}\]

\[(\mathcal{M}(f_1 \ast f_2))(s) = (\mathcal{M} f_1)(s) \cdot (\mathcal{M} f_2)(s).\]
Mellin Transform Formulation: Products Random Variables

**Theorem**

Let \( X_i \)'s independent, densities \( f_i \). \( \Xi_n = X_1 \cdots X_n \),

\[
    h_n(x_n) = (f_1 \ast \cdots \ast f_n)(x_n)
\]

\[
    (\mathcal{M} h_n)(s) = \prod_{m=1}^{n} (\mathcal{M} f_m)(s).
\]

As \( n \to \infty \), \( \Xi_n \) becomes Benford: \( Y_n = \log_B \Xi_n \),

\[
    |\text{Prob}(Y_n \mod 1 \in [a, b]) - (b - a)| \leq (b - a) \cdot \sum_{\ell \neq 0, \ell = -\infty}^{\infty} \prod_{m=1}^{n} (\mathcal{M} f_i) \left( 1 - \frac{2\pi i \ell}{\log B} \right).\]
Proof of Kossovsky’s Chain Conjecture for certain densities

Conditions

- $\{D_i(\theta)\}_{i \in I}$: one-parameter distributions, densities $f_{D_i(\theta)}$ on $[0, \infty)$. 
Proof of Kossovsky’s Chain Conjecture for certain densities

Conditions

1. \( \{D_i(\theta)\}_{i \in I} \): one-parameter distributions, densities \( f_{D_i(\theta)} \) on \([0, \infty)\).

2. \( p : \mathbb{N} \rightarrow I, \ X_1 \sim D_{p(1)}(1), \ X_m \sim D_{p(m)}(X_{m-1}). \)
Proof of Kossovsky’s Chain Conjecture for certain densities

Conditions

- \{\mathcal{D}_i(\theta)\}_{i \in I}: one-parameter distributions, densities \( f_{\mathcal{D}_i(\theta)} \) on \([0, \infty)\).
- \( p : \mathbb{N} \rightarrow I, X_1 \sim \mathcal{D}_{p(1)}(1), X_m \sim \mathcal{D}_{p(m)}(X_{m-1}) \).
- \( m \geq 2 \),

\[
f_m(x_m) = \int_0^\infty f_{\mathcal{D}_{p(m)}(1)} \left( \frac{x_m}{x_{m-1}} \right) f_{m-1}(x_{m-1}) \frac{dx_{m-1}}{x_{m-1}}
\]
Proof of Kossovsky’s Chain Conjecture for certain densities

Conditions

- $\{D_i(\theta)\}_{i \in I}$: one-parameter distributions, densities $f_{D_i(\theta)}$ on $[0, \infty)$.
- $p : \mathbb{N} \rightarrow I$, $X_1 \sim D_{p(1)}(1)$, $X_m \sim D_{p(m)}(X_{m-1})$.
- $m \geq 2$,

$$f_m(x_m) = \int_0^\infty f_{D_{p(m)}(1)} \left( \frac{x_m}{x_{m-1}} \right) f_{m-1}(x_{m-1}) \frac{dx_{m-1}}{x_{m-1}}$$

- $\lim_{n \to \infty} \sum_{\ell = -\infty}^\infty \prod_{m=1}^n \left( M f_{D_{p(m)}(1)} \right) \left( 1 - \frac{2\pi i \ell}{\log B} \right) = 0$
Proof of Kossovsky’s Chain Conjecture for certain densities

**Theorem (JKKKM)**

- *If conditions hold, as* \( n \to \infty \) *the distribution of leading digits of* \( X_n \) *tends to Benford’s law.*

- *The error is a nice function of the Mellin transforms: if* \( Y_n = \log_B X_n \), *then*

\[
|\text{Prob}(Y_n \mod 1 \in [a, b]) - (b + a)| \leq \\
(b - a) \cdot \left( \sum_{\ell = -\infty}^{\infty} \prod_{m=1}^{n} (\mathcal{M} f_{D_P(m)}(1)) \left( 1 - \frac{2\pi i \ell}{\log B} \right) \right)
\]
Example: All $X_i \sim \text{Exp}(1)$

- $X_i \sim \text{Exp}(1)$, $Y_n = \log_B \Xi_n$.
- Needed ingredients:
  - $\int_0^\infty \exp(-x)x^{s-1}dx = \Gamma(s)$.
  - $|\Gamma(1 + ix)| = \sqrt{\pi x / \sinh(\pi x)}$, $x \in \mathbb{R}$.
- $|P_n(s) - \log_{10}(s)| \leq$

$$
\log_B s \sum_{\ell=1}^\infty \left( \frac{2\pi^2 \ell / \log B}{\sinh(2\pi^2 \ell / \log B)} \right)^{n/2}.
$$
Example: All $X_i \sim \text{Exp}(1)$

**Bounds on the error**

- $|P_n(s) - \log_{10} s| \leq$
  - $3.3 \cdot 10^{-3} \log_B s$ if $n = 2$,
  - $1.9 \cdot 10^{-4} \log_B s$ if $n = 3$,
  - $1.1 \cdot 10^{-5} \log_B s$ if $n = 5$, and
  - $3.6 \cdot 10^{-13} \log_B s$ if $n = 10$.

- Error at most

$$\log_{10} s \sum_{\ell=1}^{\infty} \left( \frac{17.148\ell}{\exp(8.5726\ell)} \right)^{n/2} \leq .057^n \log_{10} s$$
Conclusions
Conclusions and Future Investigations

- See many different systems exhibit Benford behavior.
See many different systems exhibit Benford behavior.

Ingredients of proofs (logarithms, equidistribution).
Conclusions and Future Investigations

- See many different systems exhibit Benford behavior.
- Ingredients of proofs (logarithms, equidistribution).
- Applications to fraud detection / data integrity.
Conclusions and Future Investigations

- See many different systems exhibit Benford behavior.
- Ingredients of proofs (logarithms, equidistribution).
- Applications to fraud detection / data integrity.

Future work:
- Study digits of other systems.
- Develop more sophisticated tests for fraud.
References


S. J. Miller, *When the Cramér-Rao Inequality provides no information*, to appear in Communications in Information and Systems.


