Why the IRS cares about the Riemann Zeta Function and Number Theory (and why you should too!)

Steven J. Miller
sjml@williams.edu,
Steven.Miller.MC.96@aya.yale.edu

http://web.williams.edu/Mathematics/sjmiller/public_html/

Vassar College, December 10, 2019
Introduction


Interesting Question

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

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

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

Answer: Benford’s law!
Examples with First Digit Bias

Fibonacci numbers

![First 652066 Fibonacci Numbers](chart1)

Most common iPhone passcodes

![Most common iPhone passcodes](chart2)

Twitter users by # followers

![Twitter users by followers count](chart3)

Distance of stars from Earth

![Distance of stars from Earth in light years](chart4)
Summary

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

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

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

![Figure 1—Benford’s Law Distribution Leading Digit](image)

Benford’s Law (probabilities)
Background Material

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

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

- Significand: \( x = S_{10}(x) \cdot 10^k \), \( k \) integer, \( 1 \leq S_{10}(x) < 10 \).
Background Material

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

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

- \( S_{10}(x) = S_{10}(\tilde{x}) \) if and only if \( x \) and \( \tilde{x} \) have the same leading digits. Note \( \log_{10} x = \log_{10} S_{10}(x) + k \).
Background Material

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

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

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

- **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 $y = \log_{10} x \mod 1$.

Advanced: $e^{2\pi i u} = e^{2\pi i (u \mod 1)}$. 
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 \( b - a \):

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

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

- Thm: \( \beta \notin \mathbb{Q} \), \( n\beta \) 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**

\[ \{y_n\}_{n=1}^{\infty} \] 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: \( 2 = 10^{p/q} \).
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: $n\sqrt{\pi} \mod 1$ 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 $n \leq 10,000$
Logarithms and Benford’s Law

**Fundamental Equivalence**

Data set \( \{ x_i \} \) is Benford base \( B \) if \( \{ y_i \} \) is equidistributed mod 1, where \( y_i = \log_B x_i \).
Logarithms and Benford’s Law

**Fundamental Equivalence**

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

\[
x = S_{10}(x) \cdot 10^k \text{ then } \\
\log_{10} x = \log_{10} S_{10}(x) + k = \log_{10} S_{10}x \mod 1.
\]
Logarithms and Benford’s Law

**Fundamental Equivalence**

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

\[
x = S_{10}(x) \cdot 10^k \quad \text{then}
\log_{10} x = \log_{10} S_{10}(x) + k = \log_{10} S_{10}x \mod 1.
\]
Logarithms and Benford’s Law

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

Have Benford’s law $\leftrightarrow$ mantissa of logarithms of data are uniformly distributed.
The Power of the Right Perspective
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 \( a_n = r^n: \ r^{n+1} = r^n + r^{n-1} \) or \( r^2 = r + 1. \)
Fibonacci numbers are Benford base 10.

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

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

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

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

- Fibonacci numbers are Benford base 10.

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

Guess \( a_n = r^n: \quad r^{n+1} = r^n + r^{n-1} \) 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 \( a_n = r^n: \ r^{n+1} = r^n + r^{n-1} \text{ or } r^2 = r + 1. \)
  Roots \( r = \left(1 \pm \sqrt{5}\right)/2. \)
  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 \( a_n = r^n \): \( r^{n+1} = r^n + r^{n-1} \) 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 \( a_n = r^n \): \( r^{n+1} = r^n + r^{n-1} \) 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:
  \( a_{n+1} = 2a_n - a_{n-1} \)
Fibonacci numbers are Benford base 10.

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

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

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

\[ \text{take } a_0 = a_1 = 1 \text{ or } a_0 = 0, \ a_1 = 1. \]
# Digits of $2^n$

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

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$2^1$</td>
<td>1024</td>
<td>1048576</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>$2^2$</td>
<td>2048</td>
<td>2097152</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>$2^4$</td>
<td>4096</td>
<td>4194304</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>$2^8$</td>
<td>8192</td>
<td>8388608</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>$2^{16}$</td>
<td>16384</td>
<td>16777216</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>$2^{32}$</td>
<td>32768</td>
<td>33554432</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>$2^{64}$</td>
<td>65536</td>
<td>67108864</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>$2^{128}$</td>
<td>131072</td>
<td>134217728</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>$2^{256}$</td>
<td>262144</td>
<td>268435456</td>
<td>9</td>
<td>1</td>
</tr>
</tbody>
</table>
# Digits of $2^n$

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

<table>
<thead>
<tr>
<th>digit</th>
<th>#</th>
<th>Obs Prob</th>
<th>Benf Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>.300</td>
<td>.301</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>.200</td>
<td>.176</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>.100</td>
<td>.125</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>.100</td>
<td>.097</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>.100</td>
<td>.079</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>.067</td>
<td>.067</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>.033</td>
<td>.058</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>.083</td>
<td>.051</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>.017</td>
<td>.046</td>
</tr>
</tbody>
</table>
### Digits of $2^n$

First 60 values of $2^n$ (only displaying 30): $2^{10} = 1024 \approx 10^3$.

<table>
<thead>
<tr>
<th>$2^n$</th>
<th>Obs</th>
<th>Digit</th>
<th>Prob</th>
<th>Benford Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>1024</td>
<td>17</td>
<td>1</td>
<td>.300</td>
<td>.301</td>
</tr>
<tr>
<td>2048</td>
<td>12</td>
<td>2</td>
<td>.200</td>
<td>.176</td>
</tr>
<tr>
<td>4096</td>
<td>6</td>
<td>3</td>
<td>.100</td>
<td>.125</td>
</tr>
<tr>
<td>8192</td>
<td>6</td>
<td>4</td>
<td>.100</td>
<td>.097</td>
</tr>
<tr>
<td>16384</td>
<td>6</td>
<td>5</td>
<td>.100</td>
<td>.079</td>
</tr>
<tr>
<td>32768</td>
<td>4</td>
<td>6</td>
<td>.067</td>
<td>.067</td>
</tr>
<tr>
<td>65536</td>
<td>2</td>
<td>7</td>
<td>.033</td>
<td>.058</td>
</tr>
<tr>
<td>131072</td>
<td>5</td>
<td>8</td>
<td>.083</td>
<td>.051</td>
</tr>
<tr>
<td>262144</td>
<td>1</td>
<td>9</td>
<td>.017</td>
<td>.046</td>
</tr>
</tbody>
</table>

Note: The numbers in red indicate the observed frequencies, while the probabilities are calculated based on Benford's law.
## 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 (5%: $\log(\chi^2) \approx 2.74$)

$log(\chi^2)$ vs $N$ for $\pi^n$ (red) and $e^n$ (blue), $n \in \{1, \ldots, N\}$.
Logarithms and Benford’s Law: Base 10 (5%: \( \log(\chi^2) \approx 2.74 \))

\[
\log(\chi^2) \text{ vs } N \text{ for } \pi^n \text{ (red) and } e^n \text{ (blue)}, \quad n \in \{1, \ldots, N\}. \text{ Note } \pi^{175} \approx 1.0028 \cdot 10^{87}.
\]
New Result: Linear Recurrence Relations of Degree 2

- \( a_{n+1} = f(n)a_n + g(n)a_{n-1} \) with non-constant coefficients \( f(n) \) and \( g(n) \).

- Explore conditions on \( f \) and \( g \) such that the sequence generated obeys Benford’s Law for all initial values.

- First solve the closed form of the sequence \( (a_n) \), then analyze its main term.
Main idea: reduce the degree of recurrence.

\[ a_{n+1} = (\lambda(n) + \mu(n))a_n - \mu(n)\lambda(n-1)a_{n-1}, \]

and compare the coefficients:

\[ f(n) = \lambda(n) + \mu(n) \]
\[ g(n) = -\lambda(n-1)\mu(n). \]

We show that for any given pair of \( f \) and \( g \), such \( \lambda \) and \( \mu \) always exist.
Linear Recurrence Relations of Degree 2

- Recurrence relations of degree 1:
  \[ a_{n+1} = \lambda(n) a_n + b_n \]
  \[ b_n = \mu(n) b_{n-1}. \]

- \[ a_{n+1} = r(n) \left( 1 + \sum_{k=3}^{n} \prod_{i=k}^{n} \frac{\lambda(i)}{\mu(i)} + \frac{a_2}{b_1} \prod_{i=2}^{n} \frac{\lambda(i)}{\mu(i)} \right), \]
  where \( r(n) := b_1 \prod_{i=2}^{n} \mu(i). \)

- Find conditions on \( \mu, \lambda \) such that main term dominates; Benford if \( \prod \mu(i) \) is.
Examples when $f$ and $g$ are functions

- If $\mu(k) = k$, then $r(n) = n!$.

- If $\mu(k) = k^\alpha$ where $\alpha \in \mathbb{R}$, then $r(n) = (n!)^\alpha$.

- If $\mu(k) = \exp(\alpha h(k))$ where $\alpha$ is irrational and $h(k)$ is a monic polynomial, then
  \[
  \log r(n) = \alpha \sum_{k=1}^{n} h(k).
  \]

**Lemma**

The sequence $\{\alpha p(n)\}$ is equidistributed mod 1 if $\alpha \notin \mathbb{Q}$ and $p(n)$ a monic polynomial.
Examples when $f$ and $g$ are random variables

- Take $\mu(n) \sim h(n)U_n$ where the $U_n$’s are independent uniform distributions on $[0, 1]$, and $h(n)$ is a deterministic function in $n$ such that $\prod_{i=1}^{n} h(i)$ is Benford.

Then $r(n) = \prod_{i=1}^{n} h(i) \prod_{i=1}^{n} U_i$ is Benford.

- Take $\mu(n) \sim \exp(U_n)$ where the $U_n$’s are i.i.d. random variables. Then take logarithm and sum up $\log(\mu(n))$. Apply Central Limit Theorem and get a Gaussian distribution.
Use recurrence relation of degree 3 as an example. Similar main idea: reduce the degree.

Define the sequence \( \{a_n\}_{n=1}^{\infty} \) by
\[
a_{n+1} = f_1(n)a_n + f_2(n)a_{n-1} + f_3(n)a_{n-2}.
\]

Define an auxiliary sequence \( \{b_n\}_{n=1}^{\infty} \) by
\[
b_n = a_{n+1} - \lambda(n)a_n. \text{ Then } (b_n) \text{ is degree 2.}
\]
Why Benford’s Law?
Not all data sets satisfy Benford’s Law.

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

- Long street \([1, L]\): \(L = 199\) versus \(L = 999\).
- Oscillates b/w \(1/9\) and \(5/9\) with first digit 1.

Probability first digit 1 versus street length \(L\).
Not all data sets satisfy Benford’s Law.

- Oscillates b/w $1/9$ and $5/9$ with first digit 1.

Probability first digit 1 versus $\log$ (street length $L$).
Not all data sets satisfy Benford’s Law.

- Oscillates b/w $1/9$ and $5/9$ with first digit $1$.

Probability first digit $1$ versus log(street length $L$).

What if we have many streets of different lengths?
Amalgamating Streets

All houses: 1000 Streets, each from 1 to 10000.

First digit and first two digits vs Benford.
Amalgamating Streets

All houses: 1000 Streets, each from 1 to rand(10000).

First digit and first two digits vs Benford.
Amalgamating Streets

All houses: 1000 Streets, each 1 to rand(rand(10000)).

First digit and first two digits vs Benford.
Conclusion: More processes, closer to Benford.
Amalgamating Streets

All houses: 1000 Streets, each 1 to rand(rand(rand(rand(10000)))).

First digit and first two digits vs Benford.
Conclusion: More processes, closer to Benford.
Probability Review

- Let $X$ be random variable with density $p(x)$:
  - $p(x) \geq 0$; $\int_{-\infty}^{\infty} p(x)dx = 1$;
  - $\text{Prob}(a \leq X \leq b) = \int_{a}^{b} p(x)dx$.
- Mean $\mu = \int_{-\infty}^{\infty} xp(x)dx$.
- Variance $\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx$.
- Independence: knowledge of one random variable gives no knowledge of the other.
Central Limit Theorem

Normal $N(\mu, \sigma^2)$: $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$.

Theorem

If $X_1, X_2, \ldots$ independent, identically distributed random variables (mean $\mu$, variance $\sigma^2$, finite moments) then

$$S_N := \frac{X_1 + \cdots + X_N - N\mu}{\sigma \sqrt{N}}$$

converges to $N(0, 1)$. 
Central Limit Theorem: Sums of Uniform Random Variables

$X_i \sim \text{Unif}(-1/2, 1/2)$ (adjusted to mean 0, variance 1)

$Y_1 = X_1 / \sigma_{X_1}$ vs $N(0, 1)$. 
Central Limit Theorem: Sums of Uniform Random Variables

\(X_i \sim \text{Unif}(-1/2, 1/2)\) (adjusted to mean 0, variance 1)

\[Y_2 = (X_1 + X_2)/\sigma_{X_1+X_2} \text{ vs } N(0, 1).\]
Central Limit Theorem: Sums of Uniform Random Variables
\( X_i \sim \text{Unif}(−1/2, 1/2) \) (adjusted to mean 0, variance 1)

\[ Y_4 = \frac{(X_1 + X_2 + X_3 + X_4)}{\sigma_{X_1+X_2+X_3+X_4}} \text{ vs } N(0, 1). \]
Central Limit Theorem: Sums of Uniform Random Variables

\[ X_i \sim \text{Unif}(-1/2, 1/2) \text{ (adjusted to mean 0, variance 1)} \]

\[ Y_8 = \frac{(X_1 + \cdots + X_8)}{\sigma_{X_1+\cdots+X_8}} \text{ vs } \mathcal{N}(0, 1). \]
Central Limit Theorem: Sums of Uniform Random Variables

$X_i \sim \text{Unif}(−1/2, 1/2)$ (adjusted to mean 0, variance 1)

Density of $Y_4 = (X_1 + \cdots + X_4)/\sigma_{X_1+\cdots+X_4}$.

$$
\begin{align*}
\frac{1}{27} (18 + 9 \sqrt{3} \ y - \sqrt{3} \ y^3) & \quad \text{if } y = 0 \\
\frac{1}{18} (12 - 6 y^2 - \sqrt{3} \ y^3) & \quad \text{if } -\sqrt{3} < y < 0 \\
\frac{1}{54} (72 - 36 \sqrt{3} \ y + 18 y^2 - \sqrt{3} \ y^3) & \quad \text{if } \sqrt{3} < y < 2\sqrt{3} \\
\frac{1}{54} (18 \sqrt{3} \ y - 18 y^2 + \sqrt{3} \ y^3) & \quad \text{if } y = \sqrt{3} \\
\frac{1}{18} (12 - 6 y^2 + \sqrt{3} \ y^3) & \quad \text{if } 0 < y < \sqrt{3} \\
\frac{1}{54} (72 + 36 \sqrt{3} \ y + 18 y^2 + \sqrt{3} \ y^3) & \quad \text{if } -2\sqrt{3} < y \leq -\sqrt{3} \\
0 & \quad \text{if } \text{True} \\
\sqrt{3} & \quad \text{if } \text{True}
\end{align*}
$$

(Don’t even think of asking to see $Y_8$’s!)
Normal Distributions Mod 1

As $\sigma \to \infty$, $N(0, \sigma^2)$ mod 1 $\to$ Unif$(0, 1)$.

Variance is .01.
Normal Distributions Mod 1

As $\sigma \to \infty$, $N(0, \sigma^2) \mod 1 \to \text{Unif}(0, 1)$.

Variance is .1.
As $\sigma \to \infty$, $N(0, \sigma^2)$ mod 1 $\to$ Unif$(0, 1)$.

Variance is $0.5$. 
Products and Benford’s Law

Pavlovian Response: See a product, take a logarithm.
Products and Benford’s Law

Pavlovian Response: See a product, take a logarithm.

\[ X_1, X_2, \ldots \text{ nice, } W_N = X_1 \cdot X_2 \cdots X_N. \]
Products and Benford’s Law

Pavlovian Response: See a product, take a logarithm.

\[ X_1, X_2, \ldots \text{ nice}, \ W_N = X_1 \cdot X_2 \cdots X_N. \]

\[ Y_i = \log_{10} X_i, \ V_N := \log_{10} W_N. \]
Products and Benford’s Law

**Pavlovian Response:** See a product, take a logarithm.

\[ X_1, X_2, \ldots \text{ nice, } W_N = X_1 \cdot X_2 \cdots X_N. \]

\[ Y_i = \log_{10} X_i, \ V_N := \log_{10} W_N. \]

\[ V_N = \log_{10} (X_1 \cdot X_2 \cdots X_N) \]
Products and Benford’s Law

Pavlovian Response: See a product, take a logarithm.

\[ X_1, X_2, \ldots \text{ nice, } W_N = X_1 \cdot X_2 \cdots X_N. \]

\[ Y_i = \log_{10} X_i, \quad V_N := \log_{10} W_N. \]

\[ V_N = \log_{10} (X_1 \cdot X_2 \cdots X_N) \]

\[ = \log_{10} X_1 + \log_{10} X_2 + \cdots + \log_{10} X_N \]
Products and Benford’s Law

Pavlovian Response: See a product, take a logarithm.

\[ X_1, X_2, \ldots \text{ nice, } W_N = X_1 \cdot X_2 \cdot \cdots \cdot X_N. \]

\[ Y_i = \log_{10} X_i, \; V_N := \log_{10} W_N. \]

\[ V_N = \log_{10} (X_1 \cdot X_2 \cdot \cdots \cdot X_N) \]
\[ = \log_{10} X_1 + \log_{10} X_2 + \cdots + \log_{10} X_N \]
\[ = Y_1 + Y_2 + \cdots + Y_N. \]

Need distribution of \( V_N \mod 1 \), which by CLT becomes uniform, implying Benfordness!
Applications
Applications for the IRS: Detecting Fraud

A Tale of Two Steve Millers....
Detecting Fraud

Bank 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.
Can you see the cat in the tree?
Transmitting Images

How to transmit an image?

- Have an $L \times W$ grid with $LW$ pixels.

- Each pixel a triple: (Red, Green, Blue).

- Often each value in $\{0, 1, 2, 3, \ldots, 2^n - 1\}$.

- $n = 8$ gives 256 choices for each, or 16,777,216 possibilities.
Steganography

Steganography


Take one of the colors, say red, a number from 0 to 255.

Write in binary: \( r_72^7 + r_62^6 + \cdots + r_12 + r_0. \)

If change just the last or last two digits, very minor change to image.
Can you see the cat in the tree?
Can you see the cat in the tree?
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 first CDF

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

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

$X_T$ is Benford Good if there is a nice first-order CDF

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

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

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

Decay of the Fourier Transform:

$$\sum_{\ell \neq 0} |\hat{f}(T\ell)| = o(1).$$

Small translated error:

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

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

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

- **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 (Kontorovichich and M–, 2005)**

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

Theorem (Kontorovich and M–, 2005)

$X_T$ converging to $X$ as $T \to \infty$ (think spreading Gaussian). If $X_T$ is Benford good, then $X$ 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)

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

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

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

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

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

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

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

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

\[=
\hat{f}(0) \cdot (b - a) + \sum_{\ell \neq 0} \hat{f}(T\ell) \frac{e^{2\pi ib\ell} - e^{2\pi i a\ell}}{2\pi i \ell} + o(1).
\]
Riemann Zeta Function (for real part of $s$ greater than 1)

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

Geometric Series Formula: \((1 - x)^{-1} = 1 + x + x^2 + \cdots.\)

Unique Factorization: \(n = p_1^{r_1} \cdots p_m^{r_m}.\)

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

\[
= \sum_n \frac{1}{n^s}.
\]
Riemann Zeta Function

\[ |\zeta \left( \frac{1}{2} + ik \frac{4}{4} \right) |, \ k \in \{0, 1, \ldots, 65535\}. \]
3x + 1 Problem

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

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

- Conjecture: for some $n = n(x)$, $T^n(x) = 1$. 
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
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 $n = n(x)$, $T^n(x) = 1$.
- $7 \rightarrow 11$
3x + 1 Problem

- Kakutani (conspiracy), Erdös (not ready).
- \(x\) odd, \(T(x) = \frac{3x+1}{2^k}, \ 2^k \parallel 3x + 1\).
- 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, $T(x) = \frac{3x+1}{2^k}, \ 2^k \parallel 3x + 1$.

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

- $7 \rightarrow_1 11 \rightarrow_1 17 \rightarrow_2 13$
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
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 →₁ 11 →₁ 17 →₂ 13 →₃ 5 →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 $\rightarrow_1$ 11 $\rightarrow_1$ 17 $\rightarrow_2$ 13 $\rightarrow_3$ 5 $\rightarrow_4$ 1 $\rightarrow_2$ 1,
3x + 1 Problem

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

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

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

- $7 \rightarrow_1 11 \rightarrow_1 17 \rightarrow_2 13 \rightarrow_3 5 \rightarrow_4 1 \rightarrow_2 1$, 2-path (1, 1), 5-path (1, 1, 2, 3, 4).

- $m$-path: $(k_1, \ldots, k_m)$.
$3x + 1$ and Benford

**Theorem (Kontorovichich 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.
3x + 1 Data: random 10,000 digit number, $2^k \| 3x + 1$

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>
Stick Decomposition


Fixed Proportion Decomposition Process

Decomposition Process

1. Consider a stick of length $L$. 
Fixed Proportion Decomposition Process

Decomposition Process

1. Consider a stick of length $L$.

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

\[ \mathcal{L} \]

\[ p\mathcal{L} \]

\[ (1 - p)\mathcal{L} \]

\[ p^2\mathcal{L} \]

\[ p(1 - p)\mathcal{L} \]

\[ p(1 - p)\mathcal{L} \]

\[ (1 - p)^2\mathcal{L} \]
Fixed Proportion Conjecture (Joy Jing ’13)

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

(B) $p = 0.51$ and $N = 10000$. (B) $p = 0.99$ and $N = 50000$. Benford distribution overlaid.
Fixed Proportion Conjecture (Joy Jing ’13)

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

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

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

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

At $N^{th}$ level,

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

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

At \( N^{th} \) level,

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

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

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

At $N^{th}$ level,

- $2^N$ sticks
- $N+1$ distinct lengths: write $p^{N-j}(1-p)^j$ as
  
  $$p^N \left( \frac{1-p}{p} \right)^j, \quad j \in \{0, \ldots, N\}, \text{ have } \binom{N}{j} \text{ times.}$$

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

Theorem: Benford if and only if $y$ irrational.
Benford Analysis (cont)

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

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

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

where \( \text{Err}[X] \) indicates an absolute error of size at most \( X \)
Examples

\[ \rho = \frac{3}{11}, \text{ 1000 levels}; \quad y = \log_{10}(\frac{8}{3}) \notin \mathbb{Q} \]  
(irrational)
Examples

\[ p = \frac{1}{11}, \text{ 1000 levels; } y = 1 \in \mathbb{Q} \]

(rational)
Examples

\[ \rho = \frac{1}{1 + 10^{33/10}}, \text{ 1000 levels; } y = \frac{33}{10} \in \mathbb{Q} \]

(rational)
Figure: Unrestricted Decomposition: Breaking $L$ into pieces, $N = 3$. 
Conclusions and References
Conclusions and Future Investigations

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


Preliminaries

- $X_1 \cdots X_n \Leftrightarrow Y_1 + \cdots + Y_n \mod 1, \ Y_i = \log_B X_i$

- Density $Y_i$ is $g_i$, density $Y_i + Y_j$ is

\[
(g_i \ast g_j)(y) = \int_0^1 g_i(t)g_j(y - t)dt.
\]

- $h_n = g_1 \ast \cdots \ast g_n, \ \hat{h}_n(\xi) = \hat{g}_1(\xi) \cdots \hat{g}_n(\xi)$.
Modulo 1 Central Limit Theorem

**Theorem (M– and Nigrini 2007)**

\( \{Y_m\} \) independent continuous random variables on \([0, 1]\) (not necc. i.i.d.), densities \( \{g_m\} \). \( Y_1 + \cdots + Y_M \mod 1 \) converges to the uniform distribution as \( M \to \infty \) in \( L^1([0, 1]) \) if and only if for all \( n \neq 0 \), \( \lim_{M \to \infty} \hat{g}_1(n) \cdots \hat{g}_M(n) = 0 \).

◊ Gives info on rate of convergence.
Generalizations

- Levy proved for i.i.d.r.v. just one year after Benford’s paper.
- Generalized to other compact groups, with estimates on the rate of convergence.
  - Stromberg: $n$-fold convolution of a regular probability measure on a compact Hausdorff group $G$ converges to normalized Haar measure in weak-star topology iff support of the distribution not contained in a coset of a proper normal closed subgroup of $G$. 
Distribution of digits (base 10) of 1000 products $X_1 \cdots X_{1000}$, where $g_{10,m} = \phi_{11m}$.

$\phi_m(x) = m$ if $|x - 1/8| \leq 1/2m$ (0 otherwise).
Proof under stronger conditions

- Use standard CLT to show $Y_1 + \cdots + Y_M$ tends to a Gaussian.

- Use Poisson Summation to show the Gaussian tends to the uniform modulo 1.
Proof under stronger conditions

**Figure:** Plot of normal (mean 0, stdev 1).
Proof under stronger conditions

**Figure:** Plot of normal (mean 0, stdev .1) modulo 1.
Proof under stronger conditions

Figure: Plot of normal (mean 0, stdev .5) modulo 1.
Poisson Summation Formula

\[ f \text{ nice:} \]

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

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

Lemma

\[
\frac{2}{\sqrt{2\pi \sigma^2}} \int_{\sigma^{1+\delta}}^{\infty} e^{-x^2/2\sigma^2} \, dx \ll e^{-\sigma^2\delta/2}.
\]
Proof Under Weaker Conditions

Lemma

As $N \to \infty$, $p_N(x) = \frac{e^{-\pi x^2/N}}{\sqrt{N}}$ becomes equidistributed modulo 1.

\[ \int_{x=-\infty}^{\infty} p_N(x) \, dx = \frac{1}{\sqrt{N}} \sum_{n \in \mathbb{Z}} \int_{x=a}^{b} e^{-\pi (x+n)^2/N} \, dx. \]

\[ e^{-\pi (x+n)^2/N} = e^{-\pi n^2/N} + O \left( \frac{\max(1,|n|)}{N} e^{-n^2/N} \right). \]

Can restrict sum to $|n| \leq N^{5/4}$.

\[ \frac{1}{\sqrt{N}} \sum_{n \in \mathbb{Z}} e^{-\pi n^2/N} = \sum_{n \in \mathbb{Z}} e^{-\pi n^2 N}. \]
Proof Under Weaker Conditions

\[
\frac{1}{\sqrt{N}} \sum_{|n| \leq N^{5/4}} \int_{x=a}^{b} e^{-\pi(x+n)^2/N} \, dx
\]

\[
= \frac{1}{\sqrt{N}} \sum_{|n| \leq N^{5/4}} \int_{x=a}^{b} \left[ e^{-\pi n^2/N} + O \left( \frac{\max(1, |n|)}{N} e^{-n^2/N} \right) \right] \, dx
\]

\[
= \frac{b - a}{\sqrt{N}} \sum_{|n| \leq N^{5/4}} e^{-\pi n^2/N} + O \left( \frac{1}{N} \sum_{n=0}^{N^{5/4}} \frac{n + 1}{\sqrt{N}} e^{-\pi (n/\sqrt{N})^2} \right)
\]

\[
= \frac{b - a}{\sqrt{N}} \sum_{|n| \leq N^{5/4}} e^{-\pi n^2/N} + O \left( \frac{1}{N} \int_{w=0}^{N^{3/4}} (w + 1) e^{-\pi w^2} \sqrt{N} \, dw \right)
\]

\[
= \frac{b - a}{\sqrt{N}} \sum_{|n| \leq N^{5/4}} e^{-\pi n^2/N} + O \left( N^{-1/2} \right).
\]
Proof Under Weaker Conditions

Extend sums to \( n \in \mathbb{Z} \), apply Poisson Summation:

\[
\frac{1}{\sqrt{N}} \sum_{n \in \mathbb{Z}} \int_{x=a}^{b} e^{-\pi(x+n)^2/N} \, dx \approx (b - a) \cdot \sum_{n \in \mathbb{Z}} e^{-\pi n^2 N}.
\]

For \( n = 0 \) the right hand side is \( b - a \).

For all other \( n \), we trivially estimate the sum:

\[
\sum_{n \neq 0} e^{-\pi n^2 N} \leq 2 \sum_{n \geq 1} e^{-\pi nN} \leq \frac{2e^{-\pi N}}{1 - e^{-\pi N}},
\]

which is less than \( 4e^{-\pi N} \) for \( N \) sufficiently large.
Proof in General Case: Fourier input

- Fejér kernel:
  \[ F_N(x) = \sum_{n=-N}^{N} \left( 1 - \frac{|n|}{N} \right) e^{2\pi i nx}. \]

- Fejér series \( T_N f(x) \) equals
  \[ (f * F_N)(x) = \sum_{n=-N}^{N} \left( 1 - \frac{|n|}{N} \right) \hat{f}(n) e^{2\pi i nx}. \]

- Lebesgue’s Theorem: \( f \in L^1([0, 1]) \). As \( N \to \infty \), \( T_N f \) converges to \( f \) in \( L^1([0, 1]) \).
- \( T_N(f * g) = (T_Nf) * g \): convolution assoc.
Proof of Modulo 1 CLT

- Density of sum is \( h_\ell = g_1 \ast \cdots \ast g_\ell \).
- Suffices show \( \forall \epsilon : \lim_{M \to \infty} \int_0^1 |h_M(x) - 1| \, dx < \epsilon \).
- Lebesgue’s Theorem: \( N \) large,

\[
\|h_1 - T_N h_1\|_1 = \int_0^1 |h_1(x) - T_N h_1(x)| \, dx < \frac{\epsilon}{2}.
\]

- Claim: above holds for \( h_M \) for all \( M \).
Proof of Modulo 1 CLT: Proof of Claim

\[ T_N h_{M+1} = T_N (h_M * g_{M+1}) = (T_N h_M) * g_{M+1} \]

\[ \| h_{M+1} - T_N h_{M+1} \|_1 = \int_0^1 |h_{M+1}(x) - T_N h_{M+1}(x)| \, dx \]

\[ = \int_0^1 |(h_M * g_{M+1})(x) - (T_N h_M) * g_{M+1}(x)| \, dx \]

\[ = \int_0^1 \left| \int_0^1 (h_M(y) - T_N h_M(y)) g_{M+1}(x - y) \, dy \right| \, dx \]

\[ \leq \int_0^1 \int_0^1 |h_M(y) - T_N h_M(y)| g_{M+1}(x - y) \, dx \, dy \]

\[ = \int_0^1 |h_M(y) - T_N h_M(y)| dy \cdot 1 < \frac{\epsilon}{2}. \]
Proof of Modulo 1 CLT

Show $\lim_{M \to \infty} \|h_M - 1\|_1 = 0$.
Triangle inequality:

$$\|h_M - 1\|_1 \leq \|h_M - T_N h_M\|_1 + \|T_N h_M - 1\|_1.$$ 

Choices of $N$ and $\epsilon$:

$$\|h_M - T_N h_M\|_1 < \epsilon/2.$$ 

Show $\|T_N h_M - 1\|_1 < \epsilon/2.$
Proof of Modulo 1 CLT

\[ \| T_N h_M - 1 \|_1 = \int_0^1 \left| \sum_{n=-N \atop n \neq 0}^{N} \left( 1 - \frac{|n|}{N} \right) \hat{h}_M(n) e^{2\pi i n x} \right| \, dx \]

\[ \leq \sum_{n=-N \atop n \neq 0}^{N} \left( 1 - \frac{|n|}{N} \right) |\hat{h}_M(n)| \]

\[ \hat{h}_M(n) = \hat{g}_1(n) \cdots \hat{g}_M(n) \rightarrow_{M \rightarrow \infty} 0. \]

For fixed \( N \) and \( \epsilon \), choose \( M \) large so that \( |\hat{h}_M(n)| < \epsilon/4N \) whenever \( n \neq 0 \) and \( |n| \leq N \).
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 \([0, \infty)\).
- 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** \( \Xi_1 \cdot \Xi_2 \):
  \[
  \int_0^\infty f_2 \left( \frac{x}{t} \right) f_1(t) \frac{dt}{t}
  \]

\[ \therefore \text{Proof: } \text{Prob}(\Xi_1 \cdot \Xi_2 \in [0, x]): \]
\[
\int_{t=0}^\infty \text{Prob} \left( \Xi_2 \in \left[ 0, \frac{x}{t} \right] \right) f_1(t) dt
\]
\[
= \int_{t=0}^\infty F_2 \left( \frac{x}{t} \right) f_1(t) dt,
\]

differentiate.
Mellin Transform

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

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

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

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

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

**Theorem**

\( X_i \)'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 \([0, \infty)\).
- \( p : \mathbb{N} \to I, X_1 \sim D_{p(1)}(1), X_m \sim D_{p(m)}(X_{m-1}) \).
- \( m \geq 2 \),

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

\[
\lim_{n \to \infty} \sum_{\ell=-\infty}^{\infty} \prod_{m=1}^{n} (Mf_{D_{p(m)}(1)}) \left( 1 - \frac{2\pi i \ell}{\log B} \right) = 0
\]
Chains of Random Variables

Return to street problem: chain of uniforms.

Let \( D_{\text{unif}}(\theta) \) be the density of a uniform random variable on \([0, \theta]\).

Let \( X_1 \sim D_{\text{unif}}(1) \) and \( X_{n+1} \sim D_{\text{unif}}(X_n) \).
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

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

$$\log_{10} s \sum_{\ell=1}^{\infty} \left( \frac{17.148 \ell}{\exp(8.5726 \ell)} \right)^{n/2} \leq 0.057^n \log_{10} s$$