Benford’s Law and Fraud Detection, or: Why the IRS Should Care About Number Theory!

Steven J Miller
Williams College

Steven.J.Miller@williams.edu
http://www.williams.edu/go/math/sjmiller/

Bronfman Science Lunch
Williams College, October 21, 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.
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.
  - 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
Mantissa: $x = M_{10}(x) \cdot 10^k$, $k$ integer.
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.
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**

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

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

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

- 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 \( b - a \):

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

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

**Equidistribution**

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

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

- **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: $n\sqrt{\pi} \mod 1$

$n\sqrt{\pi} \mod 1$ 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 \(y_i = \log_B x_i\).
Logarithms and Benford’s Law

Fundamental Equivalence

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

- $2^n$ is Benford base 10 as $\log_{10} 2 \notin \mathbb{Q}$.
- 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: $a_n = \frac{1}{\sqrt{5}} \left( \frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left( \frac{1-\sqrt{5}}{2} \right)^n$. 
Applications
Applications for the IRS: Detecting Fraud
### Applications for the IRS: Detecting Fraud

![Image of a US Individual Income Tax Return form 1040](image-url)

- **WILLIAM J CLINTON**
- **HILLARY RODHAM CLINTON**
- **WHITE HOUSE**
- **1600 PENNSYLVANIA AVENUE N.W.**
- **WASHINGTON, DC 20500**

**Presidential Election Campaign:**
- Do you want $1 to go to the fund?
  - Yes ☑
  - No ☑

**Filing Status:**
- Married filing joint return (even if only one had income) ☑

**Exemptions:**
- Single ☑
- Chelsea ☑

**Income:**
- 1. Wages, salaries, tips, etc.: $1,400
- 2. Taxable income, interest, dividends: $6,624
- 3. Taxable income, interest, dividends: $10,000

**Adjustments to income:**
- 1. Total income: $17,624
- 2. Taxable income: $6,624
- 3. Self-employment income: $0
- 4. Self-employment deduction: $0

**Total adjusted gross income:** $24,000

**Taxes:**
- 1. Federal income tax: $3,200
- 2. State income tax: $0
- 3. Other income: $0

**Net income:** $20,800

---

This form is an example of how the IRS uses Benford's Law to detect fraud in income tax returns. By analyzing the distribution of digits in financial data, the IRS can identify patterns that deviate from the expected distribution, indicating potential fraud or errors.
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><strong>October 9, 1992</strong></td>
<td>$1,927.48</td>
</tr>
<tr>
<td></td>
<td>$27,902.31</td>
</tr>
<tr>
<td><strong>October 14, 1992</strong></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><strong>October 19, 1992</strong></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>
</tbody>
</table>

**TOTAL** $1,878,687.58
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

**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.
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
Analysis of Williams College Transactions (thanks to Richard McDowell): September 6, 2006 to June 29, 2007: 64,000+ transactions
Analysis of Williams College Transactions (thanks to Richard McDowell): September 6, 2006 to June 29, 2007: 64,000+ transactions
Analysis of Williams College Transactions (thanks to Richard McDowell): September 6, 2006 to June 29, 2007: 64,000+ transactions
Conclusions
Conclusions and Future Investigations

- Diverse systems exhibit Benford behavior.
Conclusions and Future Investigations

- Diverse systems exhibit Benford behavior.
- Ingredients of proofs (logarithms, equidistribution).
Conclusions and Future Investigations

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


