Biases: From Benford’s Law to Additive Number Theory via the IRS and Physics

Steven J Miller (Williams College)

sjm1@williams.edu
http://www.williams.edu/Mathematics/sjmiller/

Williams College, June 22, 2011
Summary

- Describe Benford’s Law and some Additive Number Theory.
- Give examples and applications.
- 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.
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
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

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

\[
\frac{\#\{n \leq N : y_n \mod 1 \in [a, b]\}}{N} \rightarrow b - a.
\]
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: \( \beta \not\in \mathbb{Q}, \ n\beta \) is equidistributed mod 1.

- Examples: \( \log_{10} 2, \log_{10} \left( \frac{1+\sqrt{5}}{2} \right) \not\in \mathbb{Q} \).
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 \( b - a \):

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

- **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: \( 2 = 10^{p/q} \).
  
  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: $n\sqrt{\pi} \mod 1$

$n\sqrt{\pi} \mod 1$ 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 \( y_i = \log_B x_i \).
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 $\log_{10} 2 \notin \mathbb{Q}$. 
Examples

- Fibonacci numbers are Benford base 10.
Fibonacci numbers are Benford base 10.

\[ a_{n+1} = a_n + a_{n-1}. \]
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 \( r^2 = r + 1 \).
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 \( r = \left(1 \pm \sqrt{5}\right)/2 \).
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 \( a_n = r^n: \ r^{n+1} = r^n + r^{n-1} \) 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: \( a_n = \frac{1}{\sqrt{5}} \left( \frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left( \frac{1-\sqrt{5}}{2} \right)^n. \)
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:
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:

\[ \diamond a_{n+1} = 2a_n \]
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: 
  \( a_{n+1} = 2a_n - a_{n-1} \)
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:
  - \( a_{n+1} = 2a_n - a_{n-1} \)
  - 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>$2^n$</th>
<th>Value</th>
<th>Digit Value</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>
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

\[ \log(\chi^2) \text{ vs } N \text{ for } \pi^n \text{ (red) and } e^n \text{ (blue)}, \]
\[ n \in \{1, \ldots, N\}. \text{ Note } \pi^{175} \approx 1.0028 \cdot 10^{87}, (5\%, \]
\[ \log(\chi^2) \approx 2.74). \]
Applications for the IRS: Detecting Fraud
Applications for the IRS: Detecting Fraud

![Image of a 1040 U.S. Individual Income Tax Return form for the year 1992, showing details of income, deductions, and other tax-related information.]
Bank Fraud

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

More Sums Than Differences
## Summary

- History of the problem.
- Examples.
- Main results.
- Describe open problems.
A finite set of integers, $|A|$ its size. Form

- **Sumset**: $A + A = \{a_i + a_j : a_i, a_j \in A\}$.
- **Difference set**: $A - A = \{a_i - a_j : a_i, a_j \in A\}$. 
A finite set of integers, $|A|$ its size. Form

- Sumset: $A + A = \{a_i + a_j : a_i, a_j \in A\}$.
- Difference set: $A - A = \{a_i - a_j : a_j, a_j \in A\}$.

**Definition**

We say $A$ is **difference dominated** if $|A - A| > |A + A|$, **balanced** if $|A - A| = |A + A|$ and **sum dominated** (or an MSTD set) if $|A + A| > |A - A|$. 
Questions

Expect **generic** set to be difference dominated:
- addition is commutative, subtraction isn’t:
- Generic pair \((x, y)\) gives 1 sum, 2 differences.
Expect generic set to be difference dominated:

- addition is commutative, subtraction isn’t:
- Generic pair \((x, y)\) gives 1 sum, 2 differences.

Questions

- Do there exist sum-dominated sets?
- If yes, how many?
Examples
Examples

- Conway: \(\{0, 2, 3, 4, 7, 11, 12, 14\}\).

- Marica (1969): \(\{0, 1, 2, 4, 7, 8, 12, 14, 15\}\).

- Freiman and Pigarev (1973): \(\{0, 1, 2, 4, 5, 9, 12, 13, 14, 16, 17, 21, 24, 25, 26, 28, 29\}\).

- Computer search: subsets of \(\{1, \ldots, 100\}\): \(\{2, 6, 7, 9, 13, 14, 16, 18, 19, 22, 23, 25, 30, 31, 33, 37, 39, 41, 42, 45, 46, 47, 48, 49, 51, 52, 54, 57, 58, 59, 61, 64, 65, 66, 67, 68, 72, 73, 74, 75, 81, 83, 84, 87, 88, 91, 93, 94, 95, 98, 100\}\).

- Recently infinite families (Hegarty, Nathanson).
Infinite Families

Key observation

If $A$ is an arithmetic progression, $|A + A| = |A - A|$. 
Infinite Families

**Key observation**

If $A$ is an arithmetic progression, 

$$|A + A| = |A - A|.$$ 

**Proof:**

- WLOG, $A = \{0, 1, \ldots, n\}$ as $A \rightarrow \alpha A + \beta$
  doesn’t change $|A + A|, |A - A|$.
Infinite Families

Key observation

If $A$ is an arithmetic progression, 
$|A + A| = |A - A|.$

Proof:

- WLOG, $A = \{0, 1, \ldots, n\}$ as $A \rightarrow \alpha A + \beta$ doesn’t change $|A + A|$, $|A - A|.$

- $A + A = \{0, \ldots, 2n\}$, $A - A = \{-n, \ldots, n\}$, both of size $2n + 1.$
Most constructions perturb an arithmetic progression.

Example:

- MSTD set $A = \{0, 2, 3, 4, 7, 11, 12, 14\}$.
- $A = \{0, 2\} \cup \{3, 7, 11\} \cup (14 - \{0, 2\}) \cup \{4\}$. 
New Construction: Notation

- \([a, b] = \{ k \in \mathbb{Z} : a \leq k \leq b \}\).

- \(A\) is a \(P_n\)-set if its sumset and its difference set contain all but the first and last \(n\) possible elements (and of course it may or may not contain some of these fringe elements).
## New Construction

### Theorem (Miller-Scheinerman ’09)

- $A = L \cup R$ be a $P_n$, MSTD set where $L \subset [1, n]$, $R \subset [n + 1, 2n]$, and $1, 2n \in A$.

- Fix a $k \geq n$ and let $m$ be arbitrary.

- $M$ any subset of $[n + k + 1, n + k + m]$ st no run of more than $k$ missing elements. Assume $n + k + 1 \notin M$.

- Set $A(M) = L \cup O_1 \cup M \cup O_2 \cup R'$, where $O_1 = [n + 1, n + k]$, $O_2 = [n + k + m + 1, n + 2k + m]$, and $R' = R + 2k + m$.

Then $A(M)$ is an MSTD set, and $\exists C > 0$ st the percentage of subsets of $\{0, \ldots, r\}$ that are in this family (and thus are MSTD sets) is at least $C/r^4$. 
Generalization: Miller-Orosz-Scheinerman

Can we find $A$ so that:

$$|\epsilon_1 A + \cdots + \epsilon_n A| > |\tilde{\epsilon}_1 A + \cdots + \tilde{\epsilon}_n A|, \quad \epsilon_i, \tilde{\epsilon}_i \in \{-1, 1\}.$$  

Consider the generalized sumset

$$f_{j_1, j_2}(A) = A + A + \cdots + A - A - A - \cdots - A,$$

where there are $j_1$ pluses and $j_2$ minuses, and set $j = j_1 + j_2$.

$P^i_n$-set

Let $A \subset [1, k]$ with $1, k, \in A$. We say $A$ is a $P^i_n$-set if any $f_{j_1, j_2}(A)$ contains all but the first $n$ and last $n$ possible elements. (Note that a $P^2_n$-set is the same as what we called a $P_n$-set earlier.)
Generalization: Miller-Orosz-Scheinerman

**Conjecture (MOS)**

For any $f_{j_1}, j_2$ and $f'_{j_1}, j'_2$, there exists a finite set of integers $A$ which is

1. a $P^j_n$-set;
2. $A \subset [1, 2n]$ and $1, 2n \in A$; and
3. $|f_{j_1}, j_2(A)| > |f'_{j_1}, j'_2(A)|$.

- Problem is finding an $A$ with $|f_{j_1}, j_2(A)| > |f'_{j_1}, j'_2(A)|$; once we find such a set, we can mirror previous construction and construct infinitely many.

- Theorem: Conjecture true for $j \in \{2, 3\}$.  

Needed input for $j = 3$: $A = \{1, 2, 5, 6, 16, 19, 22, 26, 32, 34, 35, 39, 43, 48, 49, 50\}$. Took elements in $\{2, \ldots, 49\}$ in $A$ with probability $1/3$; it took about 300000 sets to find one satisfying our conditions. To be a $P_{25}^3$-set we need to have $A + A + A \supset [n + 3, 6n - n] = [28, 125]$ and $A + A - A \supset [-n + 2, 3n - 1] = [-23, 74]$. Have $A + A + A = [3, 150]$ (all possible elements), while $A + A - A = [-48, 99] \setminus \{-34\}$ (i.e., all but -34). Thus $A$ is a $P_{25}^3$-set satisfying $|A + A + A| > |A + A - A|$, and have the needed example.

Could also take $A = \{1, 2, 3, 4, 8, 12, 18, 22, 23, 25, 26, 29, 30, 31, 32, 34, 45, 46, 49, 50\}$.
Results
Probability Review

$X$ random variable with density $f(x)$ means

- $f(x) \geq 0$;
- $\int_{-\infty}^{\infty} f(x) = 1$;
- $\text{Prob}(X \in [a, b]) = \int_{a}^{b} f(x)dx$.

Key quantities:

- Expected (Average) Value: $\mathbb{E}[X] = \int xf(x)dx$.
- Variance: $\sigma^2 = \int (x - \mathbb{E}[X])^2 f(x)dx$. 

Binomial model

Binomial model, parameter $p(n)$
Each $k \in \{0, \ldots, n\}$ is in $A$ with probability $p(n)$.

Consider uniform model ($p(n) = 1/2$):

- Let $A \in \{0, \ldots, n\}$. Most elements in $\{0, \ldots, 2n\}$ in $A + A$ and in $\{-n, \ldots, n\}$ in $A - A$.

- $\mathbb{E}[|A + A|] = 2n - 11$, $\mathbb{E}[|A - A|] = 2n - 7$. 
**Theorem**

Let $A$ be chosen from $\{0, \ldots, N\}$ according to the binomial model with constant parameter $p$ (thus $k \in A$ with probability $p$). At least $k_{SD; p} 2^{N+1}$ subsets are sum dominated.
Let $A$ be chosen from $\{0, \ldots, N\}$ according to the binomial model with constant parameter $p$ (thus $k \in A$ with probability $p$). At least $k_{\text{SD};p}2^{N+1}$ subsets are sum dominated.

$k_{\text{SD};1/2} \geq 10^{-7}$, expect about $10^{-3}$. 
Theorem

Let $A$ be chosen from $\{0, \ldots, N\}$ according to the binomial model with constant parameter $p$ (thus $k \in A$ with probability $p$). At least $k_{SD;p}2^{N+1}$ subsets are sum dominated.

- $k_{SD;1/2} \geq 10^{-7}$, expect about $10^{-3}$.

Proof ($p = 1/2$): Generically $|A| = \frac{N}{2} + O(\sqrt{N})$.

- about $\frac{N}{4} - \frac{|N-k|}{4}$ ways write $k \in A + A$.
- about $\frac{N}{4} - \frac{|k|}{4}$ ways write $k \in A - A$.
- Almost all numbers that can be in $A \pm A$ are.
- Win by controlling fringes.
Notation

- $X \sim f(N)$ means $\forall \epsilon_1, \epsilon_2 > 0, \exists N_{\epsilon_1,\epsilon_2}$ st $\forall N \geq N_{\epsilon_1,\epsilon_2}$

\[
\text{Prob} (X \not\in [(1 - \epsilon_1)f(N), (1 + \epsilon_1)f(N)]) < \epsilon_2.
\]
\[ X \sim f(N) \text{ means } \forall \epsilon_1, \epsilon_2 > 0, \exists N_{\epsilon_1, \epsilon_2} \text{ st } \forall N \geq N_{\epsilon_1, \epsilon_2} \]

\[ \text{Prob} (X \not\in [(1 - \epsilon_1)f(N), (1 + \epsilon_1)f(N)]) < \epsilon_2. \]

\[ S = |A + A|, \ D = |A - A|, \]
\[ S^c = 2N + 1 - S, \ D^c = 2N + 1 - D. \]
Notation

- $X \sim f(N)$ means $\forall \epsilon_1, \epsilon_2 > 0$, $\exists N_{\epsilon_1, \epsilon_2}$ st $\forall N \geq N_{\epsilon_1, \epsilon_2}$

  $$\text{Prob}(X \notin [(1 - \epsilon_1)f(N), (1 + \epsilon_1)f(N)]) < \epsilon_2.$$  

- $S = |A + A|$, $D = |A - A|$
  
  $S^c = 2N + 1 - S$, $D^c = 2N + 1 - D$.

New model: Binomial with parameter $p(N)$:

- $1/N = o(p(N))$ and $p(N) = o(1)$;
- $\text{Prob}(k \in A) = p(N)$.

Conjecture (Martin-O’Bryan)

As $N \to \infty$, $A$ is a.s. difference dominated.
Main Result

**Theorem (Hegarty-Miller)**

\[ p(N) \text{ as above, } g(x) = 2 \frac{e^{-x} - (1-x)}{x}. \]

- \[ p(N) = o(N^{-1/2}): D \sim 2S \sim (Np(N))^2; \]
- \[ p(N) = cN^{-1/2}: D \sim g(c^2)N, S \sim g \left( \frac{c^2}{2} \right) N \]
  \( (c \to 0, D/S \to 2; c \to \infty, D/S \to 1); \)
- \[ N^{-1/2} = o(p(N)): S^c \sim 2D^c \sim 4/p(N)^2. \]

Can generalize to binary linear forms, still have critical threshold.
Key input: recent strong concentration results of Kim and Vu (Applications: combinatorial number theory, random graphs, ...).
Key input: recent strong concentration results of Kim and Vu (Applications: combinatorial number theory, random graphs, ...).

Example (Chernoff): $t_i$ iid binary random variables, $Y = \sum_{i=1}^{n} t_i$, then

$$\forall \lambda > 0 : \text{Prob} \left( \left| Y - \mathbb{E}[Y] \right| \geq \sqrt{\lambda n} \right) \leq 2e^{-\lambda/2}.$$
Key input: recent strong concentration results of Kim and Vu (Applications: combinatorial number theory, random graphs, ...).

Example (Chernoff): $t_i$ iid binary random variables, $Y = \sum_{i=1}^{n} t_i$, then

$$\forall \lambda > 0 : \text{Prob} \left( |Y - \mathbb{E}[Y]| \geq \sqrt{\lambda n} \right) \leq 2e^{-\lambda/2}.$$ 

Need to allow dependent random variables.
Key input: recent strong concentration results of Kim and Vu (Applications: combinatorial number theory, random graphs, ...).

Example (Chernoff): $t_i$ iid binary random variables, $Y = \sum_{i=1}^{n} t_i$, then

$$\forall \lambda > 0 : \operatorname{Prob} \left( \left| Y - \mathbb{E}[Y] \right| \geq \sqrt{\lambda n} \right) \leq 2e^{-\lambda/2}.$$

Need to allow dependent random variables.

Sketch of proofs: $X \in \{S, D, S^c, D^c\}$.

1. Prove $\mathbb{E}[X]$ behaves asymptotically as claimed;
2. Prove $X$ is strongly concentrated about mean.
Transition Behavior

**Figure:** Plot of $|A - A|/|A + A|$ for ten $A$ chosen uniformly from \{1, \ldots, n\} ($n = 10,000$ on the left and $100,000$ on the right) with probability $p(n) = c/\sqrt{n}$ versus $g(c^2)/g(c^2/2)$. 
Transition Behavior

Figure: Plot of $|A - A|/|A + A|$ for ten $A$ chosen uniformly from $\{1, \ldots, n\}$ with probability $p(n) = c/\sqrt{n}$ ($n = 1,000,000$) versus $g(c^2)/g(c^2/2)$ (second plot is just a zoom in of the first).
To further investigate the transition behavior, we fixed two values of $c$ and studied the ratio for various $n$. We chose $c = .01$ (where the ratio should converge to 1.99997) and $c = .1$ (where the ratio should converge to 1.99667); the results are displayed in Table 1.

<table>
<thead>
<tr>
<th>$n$</th>
<th>Observed Ratio ($c = .01$)</th>
<th>Observed Ratio ($c = .1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100,000</td>
<td>1.123</td>
<td>1.873</td>
</tr>
<tr>
<td>1,000,000</td>
<td>1.614</td>
<td>1.956</td>
</tr>
<tr>
<td>10,000,000</td>
<td>1.871</td>
<td>1.984</td>
</tr>
<tr>
<td>100,000,000</td>
<td>1.960</td>
<td>1.993</td>
</tr>
</tbody>
</table>

Table: Observed ratios of $|A - A|/|A + A|$ for $A$ chosen with the binomial model $p(n) = cn^{-1/2}$ for $k \in \{0, \ldots, n - 1\}$ for $c = .01$ and .1; as $n \to \infty$ the ratios should respectively converge to 1.99997 and 1.99667. Each observed data point is the average from 10 randomly chosen $A$'s, except the last one for $c = .1$ which was for just one randomly chosen $A$. 
Open Questions

- Is there a set $A$ such that $A$ and $A + A$ are MSTD sets?

- Do a positive percentage of sets $A$ have $A + A$ sum-dominant?

- For linear combinations of sums / differences, is each ordering possible? IE,
  $$|A + A + A + A| > |A + A - A - A| > |A + A + A - A|?$$

- Can one give explicit constructions of large families of such sets?
Open Questions

- Is there a set $A$ such that $A$ and $A + A$ are MSTD sets?

- Do a positive percentage of sets $A$ have $A + A$ sum-dominant? YES!

- For linear combinations of sums / differences, is each ordering possible? IE,
  
  $|A + A + A + A| > |A + A - A - A| > |A + A + A - A|$?

- Can one give explicit constructions of large families of such sets?