

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/sjmillier/`

Williams College, June 22, 2011

Summary

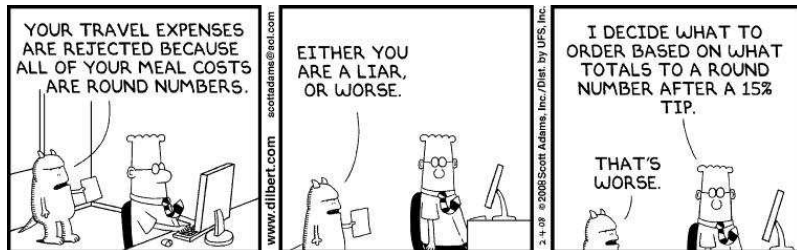
- Describe Benford's Law and some Additive Number Theory.
- Give examples and applications.
- Describe open problems.

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: $\log_{10}(x) = \log_{10}(\tilde{x}) \pmod{1}$ 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

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

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

Equidistribution and Benford's Law

Equidistribution

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

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

- 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 $y_n \bmod 1 \in [a, b]$ tends to $b - a$:

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

- 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

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

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

- 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

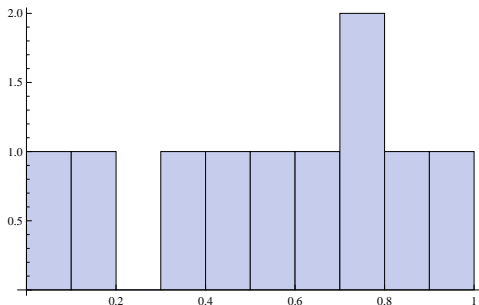
Equidistribution

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

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

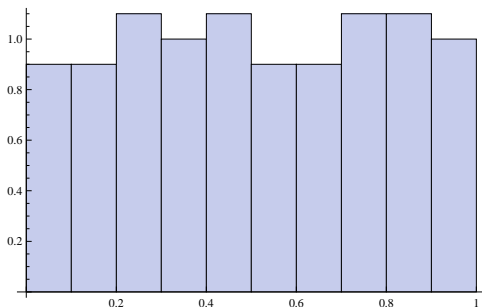
- Thm: $\beta \notin \mathbb{Q}$, $n\beta$ is equidistributed mod 1.
- Examples: $\log_{10} 2, \log_{10} \left(\frac{1+\sqrt{5}}{2}\right) \notin \mathbb{Q}$.
Proof: if rational: $2 = 10^{p/q}$.
 Thus $2^q = 10^p$ or $2^{q-p} = 5^p$, impossible.

Example of Equidistribution: $n\sqrt{\pi} \bmod 1$



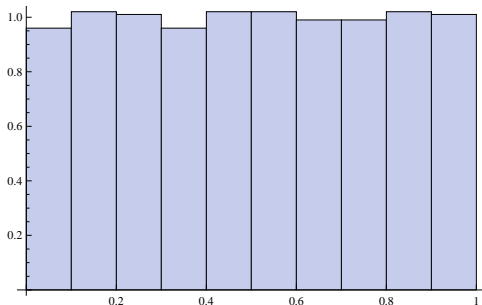
$n\sqrt{\pi} \bmod 1$ for $n \leq 10$

Example of Equidistribution: $n\sqrt{\pi} \bmod 1$



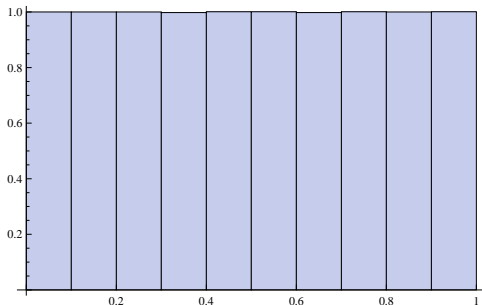
$n\sqrt{\pi} \bmod 1$ for $n \leq 100$

Example of Equidistribution: $n\sqrt{\pi} \bmod 1$



$n\sqrt{\pi} \bmod 1$ for $n \leq 1000$

Example of Equidistribution: $n\sqrt{\pi} \bmod 1$



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

Logarithms and Benford's Law

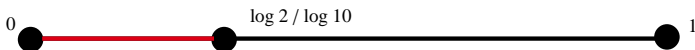
Fundamental Equivalence

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

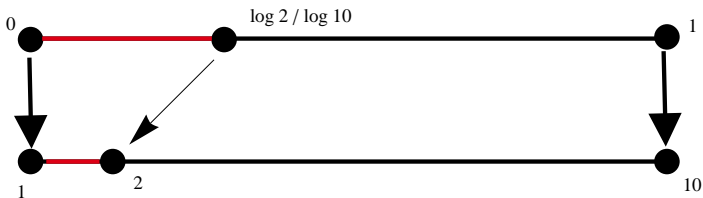
Data set $\{x_i\}$ is Benford base B if $\{y_i\}$ is equidistributed mod 1, where $y_i = \log_B x_i$.



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 $\log_{10} 2 \notin \mathbb{Q}$.

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 $r^2 = r + 1$.

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

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 $a_n = r^n$: $r^{n+1} = r^n + r^{n-1}$ 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 $a_n = r^n$: $r^{n+1} = r^n + r^{n-1}$ 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 $a_n = r^n$: $r^{n+1} = r^n + r^{n-1}$ 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:**

$$\diamond a_{n+1} = 2a_n - a_{n-1}$$

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)

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

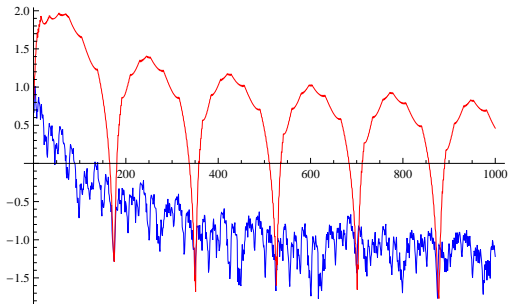
Logarithms and Benford's Law

χ^2 values for α^n , $1 \leq n \leq N$ (5% 15.5).

N	$\chi^2(\gamma)$	$\chi^2(e)$	$\chi^2(\pi)$
100	0.72	0.30	46.65
200	0.24	0.30	8.58
400	0.14	0.10	10.55
500	0.08	0.07	2.69
700	0.19	0.04	0.05
800	0.04	0.03	6.19
900	0.09	0.09	1.71
1000	0.02	0.06	2.90

Logarithms and Benford's Law: Base 10

$\log(\chi^2)$ vs N for π^n (red) and e^n (blue),
 $n \in \{1, \dots, N\}$. Note $\pi^{175} \approx 1.0028 \cdot 10^{87}$, (5%,
 $\log(\chi^2) \approx 2.74$).



Applications

Applications for the IRS: Detecting Fraud

Department of the Treasury - Internal Revenue Service
1040 U.S. Individual Income Tax Return 1989

For the year **1989** or other tax year beginning **1989**, ending **1989** (OMB No. 1545-0047)

WILLIAM J. Last name
CLINTON Last name
429-92-9947 Your social security number
HILJARY First name
BODHAM Last name
353-40-2536 Spouse's social security no.

1800 CENTER Street address (Include apt. no. if P.O. box, see page 1)
31701 E. ROCK City or town, state, and ZIP code (If a foreign address, see page 1)

ARKANSAS State
72206 ZIP code

CLIN Presidential Election
Company Do you want \$1 to go to this fund? Yes No
 Do you want \$1 to go to this fund? Yes No

Filing Status
 1 Single
 2 Married filing joint return (even if only one had income)
 3 Married filing separate returns. Enter spouse's social security number above and full name here.
 4 Head of household (must be qualifying person). (See page 7 of instructions.) If the qualifying person is your child but not your dependent, enter child's name here.
 5 Qualifying widow(er) with dependent child (your spouse died in 1981). (See page 7 of instructions.)

Exemptions
 6a Self
 b Spouse
 c Dependents: (If you are a dependent, enter "1" in column 2.)
 01 Spouse
 02 Child
 03 Other relative
 04 Other person
 05 Other person
 06 Other person
 07 Other person
 08 Other person
 09 Other person
 10 Other person
 11 Other person
 12 Other person
 13 Other person
 14 Other person
 15 Other person
 16 Other person
 17 Other person
 18 Other person
 19 Other person
 20 Other person
 21 Other person
 22 Other person
 23 Other person
 24 Other person
 25 Other person
 26 Other person
 27 Other person
 28 Other person
 29 Other person
 30 Other person
 31 Other person
 32 Other person
 33 Other person
 34 Other person
 35 Other person
 36 Other person
 37 Other person
 38 Other person
 39 Other person
 40 Other person
 41 Other person
 42 Other person
 43 Other person
 44 Other person
 45 Other person
 46 Other person
 47 Other person
 48 Other person
 49 Other person
 50 Other person
 51 Other person
 52 Other person
 53 Other person
 54 Other person
 55 Other person
 56 Other person
 57 Other person
 58 Other person
 59 Other person
 60 Other person
 61 Other person
 62 Other person
 63 Other person
 64 Other person
 65 Other person
 66 Other person
 67 Other person
 68 Other person
 69 Other person
 70 Other person
 71 Other person
 72 Other person
 73 Other person
 74 Other person
 75 Other person
 76 Other person
 77 Other person
 78 Other person
 79 Other person
 80 Other person
 81 Other person
 82 Other person
 83 Other person
 84 Other person
 85 Other person
 86 Other person
 87 Other person
 88 Other person
 89 Other person
 90 Other person
 91 Other person
 92 Other person
 93 Other person
 94 Other person
 95 Other person
 96 Other person
 97 Other person
 98 Other person
 99 Other person
 00 Other person

CHITSEA Name of dependent
431-43-0195 Social security number
DAUGHTER Relationship to you
22 No. of other dependents on this return
1 No. of other dependents on this return

1 Total number of exemptions claimed
7 **346,444**

Income
 7 Wages, salaries, tips, etc. (attach Form W-2) **586,510**
 8a Taxable interest income (attach Schedule D if any) **181**
 8b Tax-exempt interest income. (Don't include on line 8b) **3,181**
 9 Dividend income (attach Schedule D if over \$400) **3**
 10 Taxable refunds of state and local income taxes, if any, from worksheet on page 11 of instructions **10**
 11 Annuity received **11**
 12 Business income or loss (attach Schedule C) **12**
 13 Capital gain or loss (attach Schedule D) **14**
 14 Capital gain distributions not reported on line 13 **15**
 15 Other gains or losses (attach Form 4797) **16**
 16a Total IRA distributions **16a**
 16b Rollover from another IRA **16b**
 17a Total pensions and annuities **17a**
 17b Rollover from another pension plan **17b**
 18 Pensions, royalties, partnerships, estates, trusts, etc. (attach Schedule F) **18**
 19 Farm income or loss (attach Schedule F) **19**
 20 Unemployment compensation (attach Form 1042-S) **20**
 21a Social security benefits **21a**
 21b Social Security benefits received from another country **21b**
 22 Other income (list type and amount) **22**
 23 Add the amounts from lines 7 through 22. This is your total income **23**

Adjustments to Income
 24 Your IRA deduction, from applicable worksheet on page 14 or 15 **24**
 25 Spouse's IRA deduction, from applicable worksheet on page 14 or 15 **25**
 26 Self-employed health insurance deduction, from worksheet on page 16 **26**
 27 Raffle, retirement plan, and self-employed SEP, SIMPLE, and qualified plan deductions **27**
 28 Penalty on early withdrawal of savings **28**
 29 Alimony paid (see instructions) **29**
 30 Add the amounts from lines 24 through 29 **30**

Adjusted Gross Income
 31 Subtract line 30 from line 23. This is your adjusted gross income. If you file a line item 112 and you have a credit for state and local income taxes, see "State and Local Income Taxes" (line 112) on page 20 of the instructions. If you need 483 to figure your net, see page 10 of the instructions. **31**

Gross Income
 32 **346,168**

Handwritten notes:
 "Do you have family negative feedback?"
 "How is it?"
 "not entered"

Detecting Fraud

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.

Statement

A finite set of integers, $|A|$ its size. Form

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

Statement

A finite set of integers, $|A|$ its size. Form

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

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.

Questions

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 $\{1, \dots, 100\}$: $\{2, 6, 7, 9, 13, 14, 16, 18, 19, 22, 23, 25, 30, 31, 33, 37, 39, 41, 42, 45, 46, 47, 48, 49, 51, 52, 54, 57, 58, 59, 61, 64, 65, 66, 67, 68, 72, 73, 74, 75, 81, 83, 84, 87, 88, 91, 93, 94, 95, 98, 100\}$.
- 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, $A = \{0, 1, \dots, n\}$ 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, $A = \{0, 1, \dots, n\}$ as $A \rightarrow \alpha A + \beta$ doesn't change $|A + A|, |A - A|$.
- $A + A = \{0, \dots, 2n\}$, $A - A = \{-n, \dots, n\}$, both of size $2n + 1$. □

Previous Constructions

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 $\{0, \dots, r\}$ 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:

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

Consider the generalized sumset

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

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

P_n^j -set

Let $A \subset [1, k]$ with $1, k, \in A$. We say A is a P_n^j -set if any $f_{j_1, j_2}(A)$ contains all but the first n and last n possible elements. (Note that a P_n^2 -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_n^j -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 $j \in \{2, 3\}$.

Proof of Generalization

- 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 $\{2, \dots, 49\}$ 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 $k \in \{0, \dots, n\}$ is in A with probability $p(n)$.

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

- Let $A \in \{0, \dots, n\}$. Most elements in $\{0, \dots, 2n\}$ in $A + A$ and in $\{-n, \dots, n\}$ in $A - A$.
- $\mathbb{E}[|A + A|] = 2n - 11$, $\mathbb{E}[|A - A|] = 2n - 7$.

Martin and O'Bryant '06

Theorem

Let A be chosen from $\{0, \dots, N\}$ 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.

Martin and O'Bryant '06

Theorem

Let A be chosen from $\{0, \dots, N\}$ 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} .

Martin and O'Bryant '06

Theorem

Let A be chosen from $\{0, \dots, N\}$ 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} .
- 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 $\forall N \geq N_{\epsilon_1, \epsilon_2}$

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

Notation

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

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

- $\mathcal{S} = |A + A|, \mathcal{D} = |A - A|,$
 $\mathcal{S}^c = 2N + 1 - \mathcal{S}, \mathcal{D}^c = 2N + 1 - \mathcal{D}.$

Notation

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

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

- $\mathcal{S} = |A + A|, \mathcal{D} = |A - A|,$
 $\mathcal{S}^c = 2N + 1 - \mathcal{S}, \mathcal{D}^c = 2N + 1 - \mathcal{D}.$

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'Bryant)

As $N \rightarrow \infty$, A is a.s. difference dominated.

Main Result

Theorem (Hegarty-Miller)

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

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

Can generalize to binary linear forms, still have **critical threshold**.

Inputs

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

Inputs

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

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

Inputs

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.

Inputs

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.

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

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

Transition Behavior

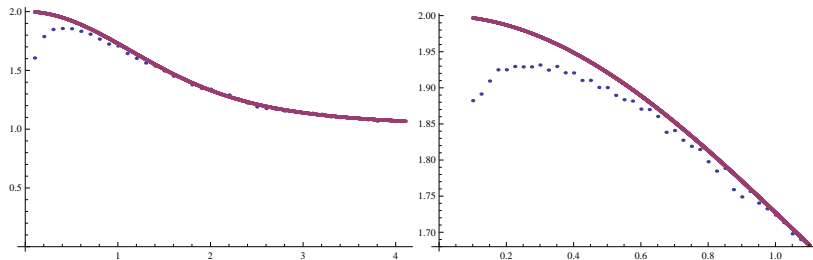


Figure: Plot of $|A - A|/|A + A|$ for ten A chosen uniformly from $\{1, \dots, n\}$ ($n = 10,000$ on the left and $100,000$ on the right) with probability $p(n) = c/\sqrt{n}$ versus $g(c^2)/g(c^2/2)$.

Transition Behavior

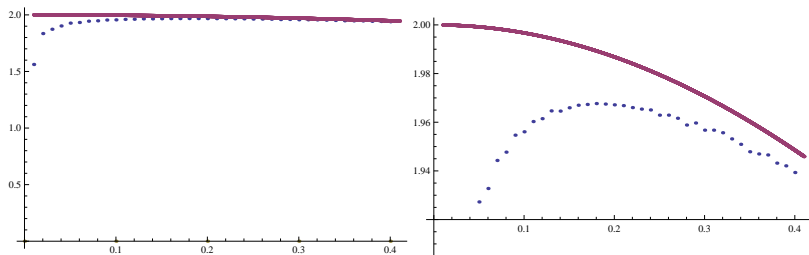


Figure: Plot of $|A - A|/|A + A|$ for ten A chosen uniformly from $\{1, \dots, n\}$ with probability $p(n) = c/\sqrt{n}$ ($n = 1,000,000$) versus $g(c^2)/g(c^2/2)$ (second plot is just a zoom in of the first).

Transition Behavior (cont)

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.

n	Observed Ratio ($c = .01$)	Observed Ratio ($c = .1$)
100,000	1.123	1.873
1,000,000	1.614	1.956
10,000,000	1.871	1.984
100,000,000	1.960	1.993

Table: Observed ratios of $|A - A|/|A + A|$ for A chosen with the binomial model $p(n) = cn^{-1/2}$ for $k \in \{0, \dots, n-1\}$ for $c = .01$ and $.1$; as $n \rightarrow \infty$ 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?