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

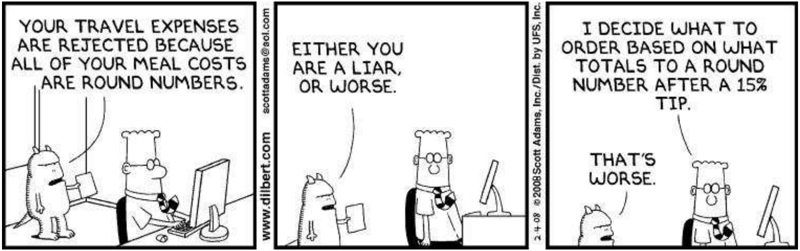
- Describe Benford's Law and some Additive Number Theory.
- Give examples and applications.
- Describe open problems.

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.

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

Key observation: $\log_{10}(x) = \log_{10}(\widetilde{x}) \bmod 1$ if and only if x and $\widetilde{x}$ 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 $b - a$:

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

Equidistribution and Benford's Law

Equidistribution

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

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

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

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

- 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 $b - a$:

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

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

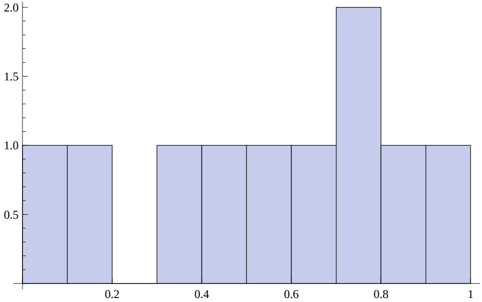
$$\frac{\#\{n \le N : y_n \bmod 1 \in [a,b]\}}{N} \to b-a.$$

- 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: $2 = 10^{p/q}$.

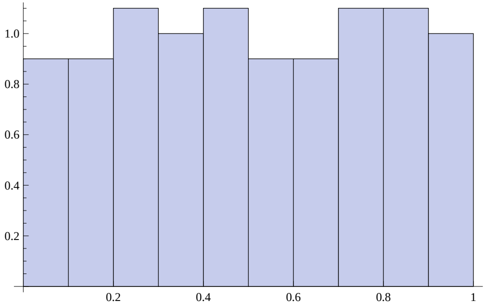
  Thus $2^q = 10^p$ or $2^{q-p} = 5^p$, impossible.

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



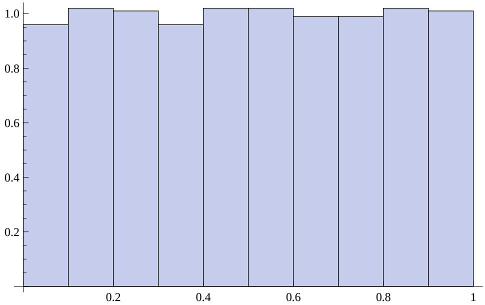
$n\sqrt{\pi} \bmod 1$ for $n \leq 10$

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



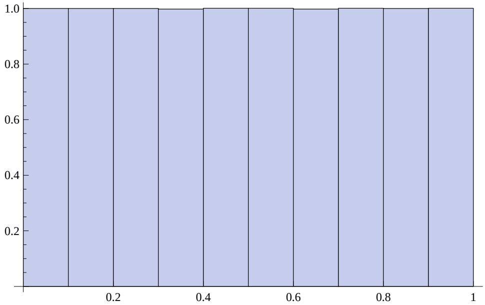
$n\sqrt{\pi}$ mod 1 for $n \leq 100$

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



$n\sqrt{\pi} \bmod 1$ 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

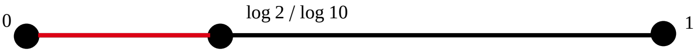
Fundamental Equivalence

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

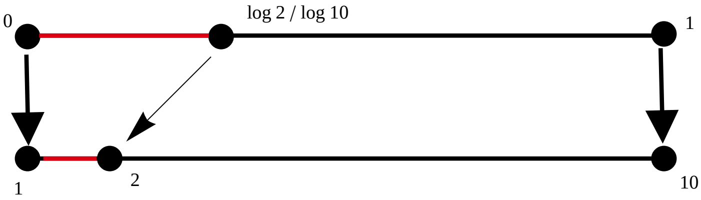
Data set $\{x_i\}$ is Benford base B if $\{y_i\}$ is equidistributed mod 1, where $y_i = \log_B x_i$.



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:
  ◇ $a_{n+1} = 2a_n$

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)

			digit	#	Obs Prob	Benf Prob
1	1024	1048576	digit	#	Obs Prob	Benf Prob
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

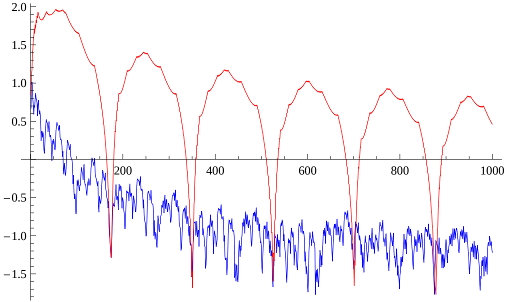
Logarithms and Benford's Law

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

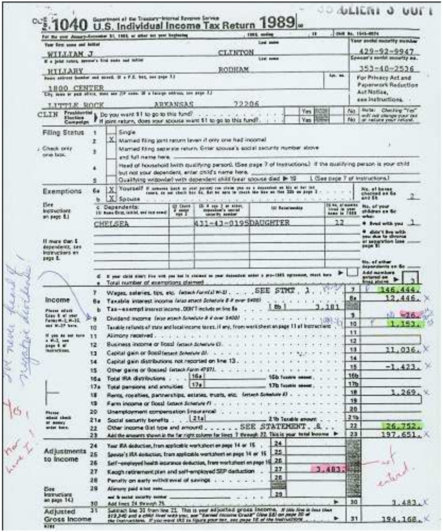
Logarithms and Benford's Law: Base 10

$\log(\chi^2)$ 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


CLIENT'S COPY

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

WILLIAM J CLINTON 429-92-9947
HILLARY RODHAM 353-40-2536
1800 CENTER
LITTLE ROCK ARKANSAS 72206

Presidential Election Campaign: Do you want $1 to go to this fund? If joint return, does your spouse want $1 to go to this fund?

Filing Status
1 Single
2 X Married filing joint return (even if only one had income)
3 Married filing separate return. Enter spouse's social security number above and full name here.
4 Head of household (with 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 (year spouse died ▶ 19). (See page 7 of Instructions.)

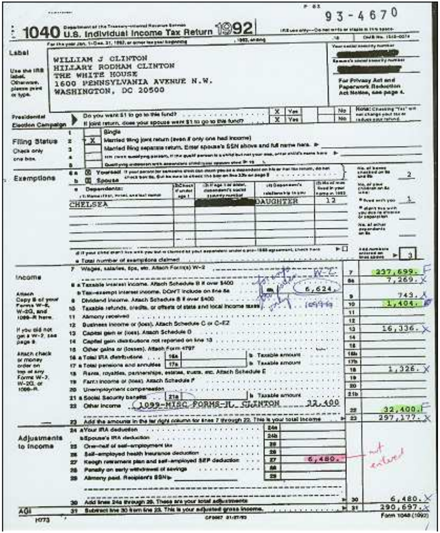
Exemptions
6a X Yourself
b X Spouse
c Dependents: CHELSEA 431-43-0195 DAUGHTER 12
e Total number of exemptions claimed 3

Income
7 Wages, salaries, tips, etc. (attach Form(s) W-2) SEE STMT 1 — 146,444.
8a Taxable interest income — 12,446.
b Tax-exempt interest income. DON'T include on line 8a — 3,181
9 Dividend income — 26.
10 Taxable refunds of state and local income taxes, if any, from worksheet on page 11 of Instructions — 1,153.
11 Alimony received
12 Business income or (loss) (attach Schedule C)
13 Capital gain or (loss) (attach Schedule D) — 11,036.
14 Capital gain distributions not reported on line 13
15 Other gains or (losses) (attach Form 4797) — -1,423.
16a Total IRA distributions; 16b Taxable amount
17a Total pensions and annuities; 17b Taxable amount
18 Rents, royalties, partnerships, estates, trusts, etc. (attach Schedule E) — 1,269.
19 Farm income or (loss) (attach Schedule F)
20 Unemployment compensation (insurance)
21a Social security benefits; 21b Taxable amount
22 Other income (list type and amount) SEE STATEMENT 8 — 26,752.
23 Add the amounts shown in the far right column for lines 7 through 22. This is your total income — 197,651.

Adjustments to Income
24 Your IRA deduction, from applicable worksheet on page 14 or 15
25 Spouse's IRA deduction, from applicable worksheet on page 14 or 15
26 Self-employed health insurance deduction, from worksheet on page 15
27 Keogh retirement plan and self-employed SEP deduction — 3,483.
28 Penalty on early withdrawal of savings
29 Alimony paid
30 Add lines 24 through 29 — 3,483.

Adjusted Gross Income
31 Subtract line 30 from line 23. This is your adjusted gross income. — 194,168.

Applications for the IRS: Detecting Fraud


93-4670

1040 Department of the Treasury—Internal Revenue Service
U.S. Individual Income Tax Return 1992

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

Your social security number [redacted]
Spouse's social security number [redacted]

For Privacy Act and Paperwork Reduction Act Notice, see page 4.

Presidential Election Campaign
Do you want $1 to go to this fund? X Yes
If joint return, does your spouse want $1 to go to this fund? X Yes

Filing Status
1 Single
2 X Married filing joint return (even if only one had income)
3 Married filing separate return. Enter spouse's SSN above and full name here.

Exemptions
6a X Yourself
b X Spouse
c Dependents:
CHELSEA [redacted] DAUGHTER 12

e Total number of exemptions claimed 3

Income
7 Wages, salaries, tips, etc. Attach Form(s) W-2 — 237,699.
8a Taxable interest income. Attach Schedule B if over $400 — 7,269.
b Tax-exempt interest income. DON'T include on line 8a — 6,624
9 Dividend income. Attach Schedule B if over $400 — 743.
10 Taxable refunds, credits, or offsets of state and local income taxes — 1,404.
11 Alimony received
12 Business income or (loss). Attach Schedule C or C-EZ
13 Capital gain or (loss). Attach Schedule D — 16,336.
14 Capital gain distributions not reported on line 13
15 Other gains or (losses). Attach Form 4797
16a Total IRA distributions
17a Total pensions and annuities
18 Rents, royalties, partnerships, estates, trusts, etc. Attach Schedule E — 1,326.
19 Farm income or (loss). Attach Schedule F
20 Unemployment compensation
21a Social Security benefits
22 Other income 1099-MISC FORMS-H. CLINTON 32,400 — 32,400.
23 Add the amounts in the far right column for lines 7 through 22. This is your total income — 297,177.

Adjustments to Income
24a Your IRA deduction
b Spouse's IRA deduction
25 One-half of self-employment tax
26 Self-employed health insurance deduction
27 Keogh retirement plan and self-employed SEP deduction — 6,480.
28 Penalty on early withdrawal of savings
29 Alimony paid.
30 Add lines 24a through 29. These are your total adjustments — 6,480.

AGI
31 Subtract line 30 from line 23. This is your adjusted gross income — 290,697.

Form 1040 (1992)

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, $A = \{0, 1, \ldots, n\}$ as $A \to \alpha A + \beta$ 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 $A \to \alpha A + \beta$ doesn't change $|A + A|$, $|A - A|$.
- $A + A = \{0, \ldots, 2n\}$, $A - A = \{-n, \ldots, n\}$, both of size $2n + 1$. $\square$

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:

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

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_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 $A + A - A = [-48, 99] \backslash \{-34\}$ (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 $A - A$.
- $\mathbb{E}[|A + A|] = 2n - 11$, $\mathbb{E}[|A - A|] = 2n - 7$.

Martin and O'Bryant '06

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* $k_{\mathrm{SD};p} 2^{N+1}$ *subsets are sum dominated.*

Martin and O'Bryant '06

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 $k_{\mathrm{SD};p} 2^{N+1}$ subsets are sum dominated.

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

Martin and O'Bryant '06

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 $k_{\mathrm{SD};p} 2^{N+1}$ subsets are sum dominated.

- $k_{\mathrm{SD};1/2} \geq 10^{-7}$, 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 $\forall N \geq N_{\epsilon_1,\epsilon_2}$

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

Notation

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

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

Notation

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

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

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

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

Conjecture (Martin-O'Bryant)

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

Main Result

Theorem (Hegarty-Miller)

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

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

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

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

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

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

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

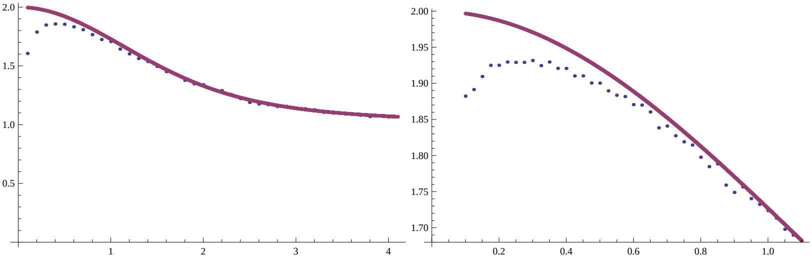


Figure: Plot of $|A-A|/|A+A|$ for ten A chosen uniformly from $\{1,\dots,n\}$ ($n = 10,000$ on the left and $100,000$ on the right) with probability $p(n) = c/\sqrt{n}$ versus $g(c^2)/g(c^2/2)$.

Transition Behavior

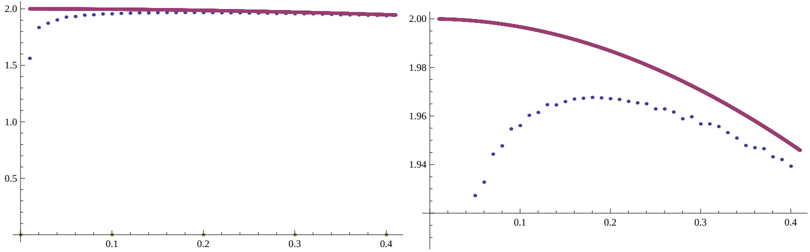


Figure: Plot of $|A - A|/|A + A|$ for ten A chosen uniformly from $\{1, \dots, n\}$ with probability $p(n) = c/\sqrt{n}$ ($n = 1,000,000$) versus $g(c^2)/g(c^2/2)$ (second plot is just a zoom in of the first).

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 $n \to \infty$ 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?