Benford's Law, Values of L-Functions and the $3x + 1$ Problem, or: Why the IRS cares about Number and Ergodic Theory

Steven J Miller
Williams College

Steven.J.Miller@williams.edu
http://web.williams.edu/Mathematics/sjmiller/public_html/

University of Illinois at Urbana-Champaign
Ergodic Theory Seminar, March 25, 2013

Introduction

Interesting Question

For a nice data set, such as the Fibonacci numbers, stock prices, street addresses of audience, ..., what percent of the leading digits are 1?

Plausible answers:

Introduction

Interesting Question

For a nice data set, such as the Fibonacci numbers, stock prices, street addresses of audience, ..., what percent of the leading digits are 1?

Plausible answers: 10%

Introduction

Interesting Question

For a nice data set, such as the Fibonacci numbers, stock prices, street addresses of audience, ..., what percent of the leading digits are 1?

Plausible answers: 10%, 11%

Introduction

Interesting Question

For a nice data set, such as the Fibonacci numbers, stock prices, street addresses of audience, ..., what percent of the leading digits are 1?

Plausible answers: 10%, 11%, about 30%.

Summary

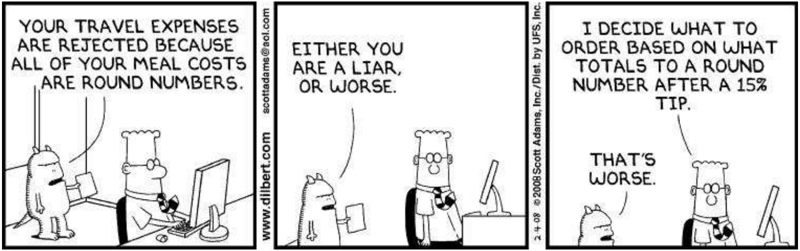
- State Benford's Law.
- Discuss examples and applications.
- Sketch proofs.
- Describe open problems.

Caveats!

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

Caveats!

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



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

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

Significands

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

$S_{10}(x) = S_{10}(\widetilde{x})$ 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.

Significands

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

$S_{10}(x) = S_{10}(\widetilde{x})$ 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 $y = \log_{10} x$.

⋄ Benefit: $e^{2\pi i (y \bmod 1)} = e^{2\pi i y}$.

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 $b - a$:

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

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 $b - a$:

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

- Thm: $\beta \notin \mathbb{Q}$, $n\beta$ is equidistributed mod 1.

Equidistribution and Benford's Law

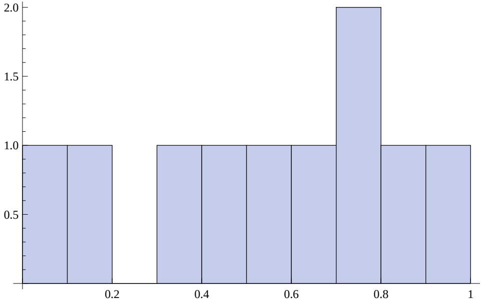
Equidistribution

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

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

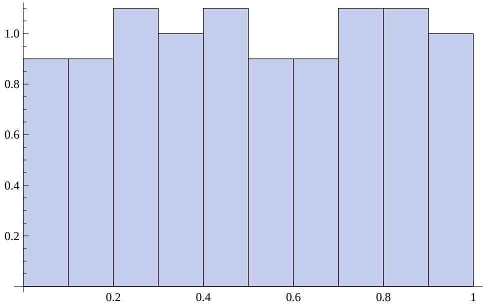
- Thm: $\beta \notin \mathbb{Q}$, $n\beta$ is equidistributed mod 1.
- Examples: $\log_{10} 2, \log_{10}\left(\frac{1+\sqrt{5}}{2}\right) \notin \mathbb{Q}$.

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



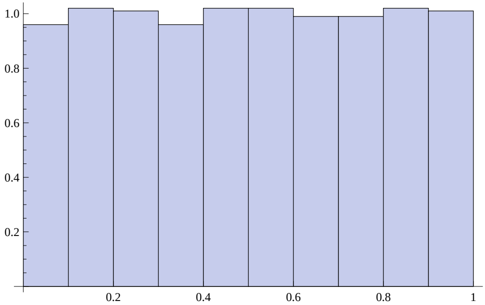
$n\sqrt{\pi}$ mod 1 for $n \le 10$

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



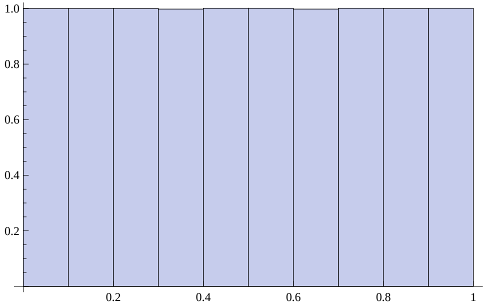
$n\sqrt{\pi}$ mod 1 for $n \leq 100$

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



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

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



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

Logarithms and Benford's Law

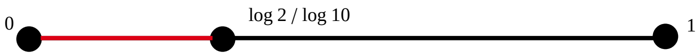
Fundamental Equivalence

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

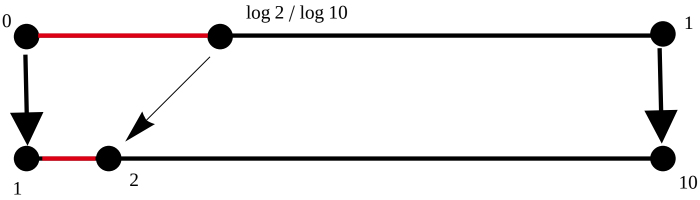
Data set $\{x_i\}$ is Benford base B if $\{y_i\}$ is equidistributed mod 1, where $y_i = \log_B x_i$.



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

- 2^n is Benford base 10 as $\log_{10} 2 \notin \mathbb{Q}$.

- Fibonacci numbers are Benford base 10.

  $a_{n+1} = a_n + a_{n-1}$.

  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

- 2^n is Benford base 10 as $\log_{10} 2 \notin \mathbb{Q}$.

- Fibonacci numbers are Benford base 10.
  $a_{n+1} = a_n + a_{n-1}$.
  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.
  $a_{n+1} = 2a_n$
  $a_{n+1} = 2a_n - a_{n-1}$
  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	1	18	.300	.301
2	2048	2097152	2	12	.200	.176
4	4096	4194304	3	6	.100	.125
8	8192	8388608	4	6	.100	.097
16	16384	16777216	5	6	.100	.079
32	32768	33554432	6	4	.067	.067
64	65536	67108864	7	2	.033	.058
128	131072	134217728	8	5	.083	.051
256	262144	268435456	9	1	.017	.046
512	524288	536870912				

Digits of 2^n

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

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

Digits of 2^n

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

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

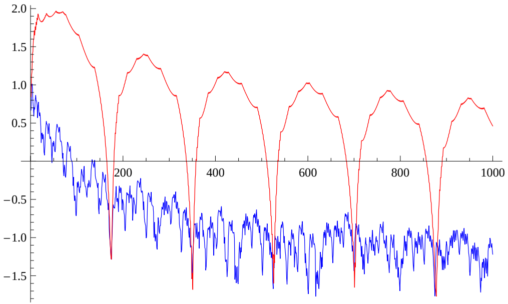
Logarithms and Benford's Law

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

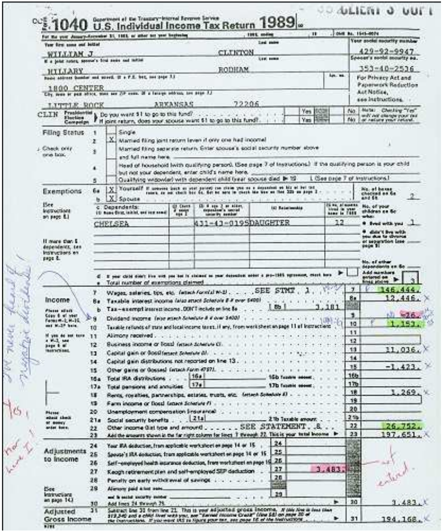
Logarithms and Benford's Law: Base 10

$\log_{10}(\chi^2)$ 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% and 8 d.f., $\log_{10}(\chi^2) \approx .44$).



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

For Privacy Act and Paperwork Reduction Act Notice, see Instructions.

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.)

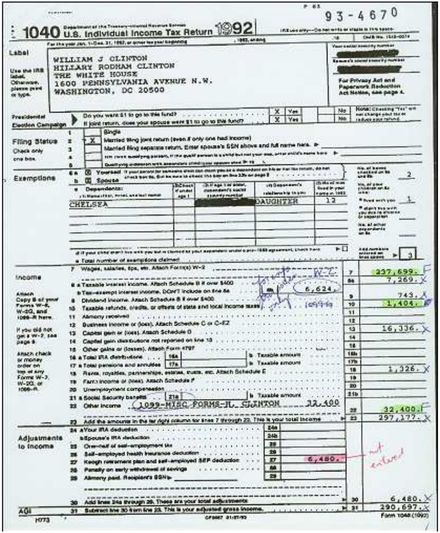
Exemptions
6a X Yourself
b X Spouse
c Dependents: CHELSEA 431-43-0195 DAUGHTER 12
e Total number of exemptions claimed 3

Income
7 Wages, salaries, tips, etc. SEE STMT 1 — 146,444.
8a Taxable interest income — 12,446.
b Tax-exempt interest income — 8b 1,101
9 Dividend income — 26.
10 Taxable refunds of state and local income taxes — 1,153.
11 Alimony received
12 Business income or (loss) — 11,036.
13 Capital gain or (loss)
14 Capital gain distributions not reported on line 13
15 Other gains or (losses) — -1,423.
16a Total IRA distributions
17a Total pensions and annuities
18 Rents, royalties, partnerships, estates, trusts, etc. — 1,269.
19 Farm income or (loss)
20 Unemployment compensation (insurance)
21a Social security benefits
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
25 Spouse's IRA deduction
26 Self-employed health insurance deduction
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 22. This is your adjusted gross income. — 194,168.

Applications for the IRS: Detecting Fraud


93-4670
1040 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
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 — 16,336.
13 Capital gain or (loss). Attach Schedule D
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 — 6,480.
26 Self-employed health insurance deduction
27 Keogh retirement plan and self-employed SEP deduction
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

Detecting Fraud

Bank Fraud

- Audit of a bank revealed huge spike of numbers starting with 4

Detecting Fraud

Bank Fraud

- Audit of a bank revealed huge spike of numbers starting with 48 and 49

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.

Benford Good Processes

Poisson Summation and Benford's Law: Definitions

- Feller, Pinkham (often exact processes)

Poisson Summation and Benford's Law: Definitions

- Feller, Pinkham (often exact processes)
- data $Y_{T,B} = \log_B \overrightarrow{X}_T$ (discrete/continuous):

$$\mathbb{P}(A) \ = \ \lim_{T\to\infty} \frac{\#\{n \in A : n \le T\}}{T}$$

Poisson Summation and Benford's Law: Definitions

- Feller, Pinkham (often exact processes)
- data $Y_{T,B} = \log_B \overrightarrow{X}_T$ (discrete/continuous):

$$\mathbb{P}(A) \ = \ \lim_{T\to\infty} \frac{\#\{n \in A : n \le T\}}{T}$$

- Poisson Summation Formula: f nice:

$$\sum_{\ell=-\infty}^{\infty} f(\ell) \ = \ \sum_{\ell=-\infty}^{\infty} \widehat{f}(\ell),$$

$$\text{Fourier transform } \widehat{f}(\xi) \ = \ \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi} dx.$$

Benford Good Process

X_T is Benford Good if there is a nice f st

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

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

Benford Good Process

X_T is Benford Good if there is a nice f st

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

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

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

Benford Good Process

X_T is Benford Good if there is a nice f st

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

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

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

Benford Good Process

X_T is Benford Good if there is a nice f st

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

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

- Small tails: $G_T(\infty) - G_T(Th(T)) = o(1)$, $G_T(-Th(T)) - G_T(-\infty) = o(1)$.
- Decay of the Fourier Transform: $\sum_{\ell \neq 0} \left| \frac{\widehat{f}(T\ell)}{\ell} \right| = o(1)$.
- Small translated error: $\mathcal{E}(a,b,T)) = \sum_{|\ell| \leq Th(T)} [E_T(b+\ell) - E_T(a+\ell)] = o(1)$.

Main Theorem

Theorem (Kontorovich and M–, 2005)

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

Main Theorem

Theorem (Kontorovich and M–, 2005)

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

- Examples
  - ◇ L-functions
  - ◇ characteristic polynomials (RMT)
  - ◇ $3x+1$ problem
  - ◇ geometric Brownian motion.

Sketch of the proof

- Structure Theorem:
  - ◇ main term is something nice spreading out
  - ◇ apply Poisson summation

Sketch of the proof

- Structure Theorem:
  - ◇ main term is something nice spreading out
  - ◇ apply Poisson summation
- Control translated errors:
  - ◇ hardest step
  - ◇ techniques problem specific

Sketch of the proof (continued)

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

Sketch of the proof (continued)

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

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

Sketch of the proof (continued)

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

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

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

Sketch of the proof (continued)

$$\sum_{\ell=-\infty}^{\infty} \mathbb{P}\left(a+\ell \le \overrightarrow{Y}_{T,B} \le b+\ell\right)$$

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

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

$$= \widehat{f}(0)\cdot(b-a) + \sum_{\ell\ne 0} \widehat{f}(T\ell)\frac{e^{2\pi i b\ell} - e^{2\pi i a\ell}}{2\pi i\ell} + o(1).$$

The $3x + 1$ Problem
and
Benford's Law

$3x+1$ Problem

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

$3x + 1$ Problem

- Kakutani (conspiracy), Erdös (not ready).
- x odd, $T(x) = \frac{3x+1}{2^k}$, $2^k || 3x+1$.
- Conjecture: for some $n = n(x)$, $T^n(x) = 1$.

$3x + 1$ Problem

- Kakutani (conspiracy), Erdös (not ready).
- x odd, $T(x) = \frac{3x+1}{2^k}$, $2^k || 3x + 1$.
- Conjecture: for some $n = n(x)$, $T^n(x) = 1$.
- 7

$3x+1$ Problem

- Kakutani (conspiracy), Erdös (not ready).
- x odd, $T(x) = \frac{3x+1}{2^k}$, $2^k || 3x+1$.
- Conjecture: for some $n = n(x)$, $T^n(x) = 1$.
- $7 \to_1 11$

$3x+1$ Problem

- Kakutani (conspiracy), Erdös (not ready).
- x odd, $T(x) = \frac{3x+1}{2^k}$, $2^k || 3x+1$.
- Conjecture: for some $n = n(x)$, $T^n(x) = 1$.
- $7 \to_1 11 \to_1 17$

$3x + 1$ Problem

- Kakutani (conspiracy), Erdös (not ready).
- x odd, $T(x) = \frac{3x+1}{2^k}$, $2^k || 3x + 1$.
- Conjecture: for some $n = n(x)$, $T^n(x) = 1$.
- $7 \to_1 11 \to_1 17 \to_2 13$

$3x + 1$ Problem

- Kakutani (conspiracy), Erdös (not ready).
- x odd, $T(x) = \frac{3x+1}{2^k}$, $2^k || 3x + 1$.
- Conjecture: for some $n = n(x)$, $T^n(x) = 1$.
- $7 \to_1 11 \to_1 17 \to_2 13 \to_3 5$

$3x + 1$ Problem

- Kakutani (conspiracy), Erdös (not ready).
- x odd, $T(x) = \frac{3x+1}{2^k}$, $2^k || 3x + 1$.
- Conjecture: for some $n = n(x)$, $T^n(x) = 1$.
- $7 \to_1 11 \to_1 17 \to_2 13 \to_3 5 \to_4 1$

$3x + 1$ Problem

- Kakutani (conspiracy), Erdös (not ready).
- x odd, $T(x) = \frac{3x+1}{2^k}$, $2^k || 3x + 1$.
- Conjecture: for some $n = n(x)$, $T^n(x) = 1$.
- $7 \to_1 11 \to_1 17 \to_2 13 \to_3 5 \to_4 1 \to_2 1$,

$3x + 1$ Problem

- Kakutani (conspiracy), Erdös (not ready).
- x odd, $T(x) = \frac{3x+1}{2^k}$, $2^k || 3x + 1$.
- Conjecture: for some $n = n(x)$, $T^n(x) = 1$.
- $7 \to_1 11 \to_1 17 \to_2 13 \to_3 5 \to_4 1 \to_2 1$,
  2-path $(1, 1)$, 5-path $(1, 1, 2, 3, 4)$.
  m-path: $(k_1, \dots, k_m)$.

Heuristic Proof of $3x + 1$ Conjecture

$$a_{n+1} = T(a_n)$$

Heuristic Proof of $3x + 1$ Conjecture

$$a_{n+1} = T(a_n)$$

$$\mathbb{E}[\log a_{n+1}] \approx \sum_{k=1}^{\infty} \frac{1}{2^k} \log\left(\frac{3a_n}{2^k}\right)$$

Heuristic Proof of $3x+1$ Conjecture

$$\begin{aligned} a_{n+1} &= T(a_n) \\ \mathbb{E}[\log a_{n+1}] &\approx \sum_{k=1}^{\infty} \frac{1}{2^k} \log\left(\frac{3a_n}{2^k}\right) \\ &= \log a_n + \log\left(\frac{3}{4}\right). \end{aligned}$$

Geometric Brownian Motion, drift $\log(3/4) < 1$.

Structure Theorem: Sinai, Kontorovich-Sinai

$$\mathbb{P}(A) = \lim_{N\to\infty} \frac{\#\{n \le N : n \equiv 1,5 \bmod 6, n \in A\}}{\#\{n \le N : n \equiv 1,5 \bmod 6\}}.$$

Structure Theorem: Sinai, Kontorovich-Sinai

$\mathbb{P}(A) = \lim_{N\to\infty} \frac{\#\{n\le N: n\equiv 1,5 \bmod 6, n\in A\}}{\#\{n\le N: n\equiv 1,5 \bmod 6\}}$.

$(k_1, \dots, k_m)$: two full arithm progressions:

$6 \cdot 2^{k_1+\cdots+k_m} p + q$.

Structure Theorem: Sinai, Kontorovich-Sinai

$$\mathbb{P}(A) = \lim_{N\to\infty} \frac{\#\{n\le N : n \equiv 1,5 \bmod 6, n\in A\}}{\#\{n \le N : n\equiv 1,5 \bmod 6\}}.$$

$(k_1, \dots, k_m)$: two full arithm progressions:

$6 \cdot 2^{k_1+\cdots+k_m} p + q$.

Theorem (Sinai, Kontorovich-Sinai)

k_i-values are i.i.d.r.v. (geometric, $1/2$):

$$\mathbb{P}\left(\frac{\log_2\left[\frac{x_m}{\left(\frac{3}{4}\right)^m x_0}\right]}{\sqrt{2m}} \le a\right) = \mathbb{P}\left(\frac{S_m - 2m}{\sqrt{2m}} \le a\right)$$

Structure Theorem: Sinai, Kontorovich-Sinai

$\mathbb{P}(A) = \lim_{N\to\infty} \frac{\#\{n\le N : n\equiv 1,5 \bmod 6, n\in A\}}{\#\{n\le N : n\equiv 1,5 \bmod 6\}}$.

$(k_1, \dots, k_m)$: two full arithm progressions:

$6 \cdot 2^{k_1+\cdots+k_m} p + q$.

Theorem (Sinai, Kontorovich-Sinai)

k_i-values are i.i.d.r.v. (geometric, $1/2$):

$$\mathbb{P}\left(\frac{\log_2\left[\frac{x_m}{\left(\frac{3}{4}\right)^m x_0}\right]}{(\log_2 B)\sqrt{2m}} \le a\right) = \mathbb{P}\left(\frac{S_m - 2m}{(\log_2 B)\sqrt{2m}} \le a\right)$$

Structure Theorem: Sinai, Kontorovich-Sinai

$\mathbb{P}(A) = \lim_{N\to\infty} \frac{\#\{n\le N: n\equiv 1,5 \bmod 6, n\in A\}}{\#\{n\le N: n\equiv 1,5 \bmod 6\}}$.

$(k_1, \dots, k_m)$: two full arithm progressions:

$6 \cdot 2^{k_1+\cdots+k_m} p + q$.

Theorem (Sinai, Kontorovich-Sinai)

k_i-values are i.i.d.r.v. (geometric, 1/2):

$$\mathbb{P}\left(\frac{\log_B\left[\frac{x_m}{\left(\frac{3}{4}\right)^m x_0}\right]}{\sqrt{2m}} \le a\right) = \mathbb{P}\left(\frac{\frac{(S_m - 2m)}{\log_2 B}}{\sqrt{2m}} \le a\right)$$

$3x + 1$ and Benford

Theorem (Kontorovich and M–, 2005)

As $m \to \infty$, $x_m/(3/4)^m x_0$ is Benford.

Theorem (Lagarias-Soundararajan 2006)

$X \geq 2^N$, for all but at most $c(B)N^{-1/36}X$ initial seeds the distribution of the first N iterates of the $3x+1$ map are within $2N^{-1/36}$ of the Benford probabilities.

Prereq: Irrationality Type

Irrationality type

α has irrationality type κ if κ is the supremum of all γ with

$$\underline{\lim}_{q\to\infty} q^{\gamma+1} \min_p \left|\alpha - \frac{p}{q}\right| = 0.$$

- Algebraic irrationals: type 1 (Roth's Thm).
- Theory of Linear Forms: $\log_B 2$ of finite type.

Prereq: Linear Forms

Theorem (Baker)

$\alpha_1, \dots, \alpha_n$ algebraic numbers height $A_j \geq 4$, $\beta_1, \dots, \beta_n \in \mathbb{Q}$ with height at most $B \geq 4$,

$$\Lambda = \beta_1 \log \alpha_1 + \cdots + \beta_n \log \alpha_n.$$

If $\Lambda \neq 0$ then $|\Lambda| > B^{-C\Omega \log \Omega'}$, with $d = [\mathbb{Q}(\alpha_i, \beta_j) : \mathbb{Q}]$, $C = (16nd)^{200n}$, $\Omega = \prod_j \log A_j$, $\Omega' = \Omega / \log A_n$.

Gives $\log_{10} 2$ of finite type, with $\kappa < 1.2 \cdot 10^{602}$:

$$|\log_{10} 2 - p/q| = |q \log 2 - p \log 10| \, / q \log 10.$$

Prereq: Quantified Equidistribution

Theorem (Erdös-Turan)

$$D_N = \frac{\sup_{[a,b]} \left|N(b-a) - \#\{n \le N : x_n \in [a,b]\}\right|}{N}$$

There is a C such that for all m:

$$D_N \le C \cdot \left(\frac{1}{m} + \sum_{h=1}^{m} \frac{1}{h} \left| \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i h x_n} \right| \right)$$

Prereq: Ideas for the proof of Erdös-Turan

Consider special case $x_n = n\alpha$, $\alpha \notin \mathbb{Q}$.

- Exponential sum $\le \frac{1}{|\sin(\pi h\alpha)|} \le \frac{1}{2||h\alpha||}$.
- Must control $\sum_{h=1}^{m} \frac{1}{h||h\alpha||}$, see irrationality type enter.
- type κ, $\sum_{h=1}^{m} \frac{1}{h||h\alpha||} = O\left(m^{\kappa-1+\epsilon}\right)$, take $m = \lfloor N^{1/\kappa} \rfloor$.

Sketch of the proof: Notation

- $[a, b]$ is an arbitrary sub-interval of $[0, 1]$.
- $c \in (0, \frac{1}{2})$ and set $M = m^c$.
- $I_\ell \ := \ \{\ell M, \ell M + 1, \dots, (\ell + 1)M - 1\}$.
- $C := \log_B 2$ an irrational number of type κ.
- $\eta(x)$ the density of the standard normal.
- S_m sum of m independent Geom(1/2) r.v.

Sketch of the proof

Step 1: Central Limit Theorem:

CLT: for any $k \in \mathbb{Z}$ have

$$\begin{aligned}\text{Prob}(C\cdot\overline{S}_m = C\cdot k) &= \text{Prob}\left(\frac{\overline{S}_m}{\sqrt{m}} = \frac{k}{\sqrt{m}}\right)\\ &= \frac{1}{\sqrt{m}}\eta\left(\frac{k}{\sqrt{m}}\right) + o\left(\frac{1}{\sqrt{m}}\right)\end{aligned}$$

Write $o\left(\frac{1}{\sqrt{m}}\right)$ as $O\left(\frac{1}{\sqrt{m}g(m)}\right)$ for some monotone increasing $g(m)$ which tends to infinity.

Sketch of the proof

Step 2: Variation in I_ℓ:

Let $k_1, k_2 \in I_\ell$. Then

$$\left|\frac{1}{\sqrt{m}}\,\eta\left(\frac{k_1}{\sqrt{m}}\right) - \frac{1}{\sqrt{m}}\,\eta\left(\frac{k_2}{\sqrt{m}}\right)\right| \leq \frac{1}{\sqrt{m}}\,e^{-\ell^2M^2/2m} \cdot \left(1 - \exp\left(-\frac{2\ell M^2 + M^2}{2m}\right)\right).$$

Means that for the ℓ we must study, there is negligible variation in the Gaussian for $k \in I_\ell$.

Sketch of the proof

Step 3: Poisson Summation:

By Poisson Summation:

$$\frac{1}{\sigma}\sum_{n=-\infty}^{\infty} e^{-n^2\pi/\sigma^2} = \sum_{n=-\infty}^{\infty} e^{-n^2\pi\sigma^2}, \quad \sigma > 0.$$

Take $\sigma^2 = \frac{2m}{\pi M^2}$, and use this to calculate the main term.

Sketch of the proof

Step 4: Quantified Equidistribution:

For any $\epsilon > 0$, letting $\delta = 1 + \epsilon - \frac{1}{\kappa} < 1$ we have

$$\#\{k \in I_\ell : kC \bmod 1 \in [a, b]\} \ = \ M(b-a) + O(M^\delta).$$

The quantification of the equidistribution of kC mod 1 is the key ingredient in proving Benford behavior base B. Follows from the Erdös-Turan Theorem.

Sketch of the proof

Step 5: Combining Pieces:

Must show as $m \to \infty$, for any $[a, b] \subset [0, 1]$,

$$P_m(a, b) \ := \ \text{Prob}(C\overline{S}_m \bmod 1 \in [a, b]) \ \longrightarrow \ b - a.$$

We have

$$\begin{aligned} P_m(a, b) &= \sum_{|\ell| \le \frac{\sqrt{m}h(m)}{M}} \text{Prob}(\overline{S}_m = k \in I_\ell : kC \bmod 1 \in [a, b]) \\ &+ \sum_{|\ell| > \frac{\sqrt{m}h(m)}{M}} \text{Prob}(\overline{S}_m = k \in I_\ell : kC \bmod 1 \in [a, b]) \\ &= \sum_{|\ell| \le \frac{\sqrt{m}h(m)}{M}} \text{Prob}(\overline{S}_m = k \in I_\ell : kC \bmod 1 \in [a, b]) + o(1). \end{aligned}$$

$3x+1$ Data: random 10,000 digit number, $2^k || 3x+1$

80,514 iterations ($(4/3)^n = a_0$ predicts 80,319); $\chi^2 = 13.5$ (5% 15.5).

Digit	Number	Observed	Benford
1	24251	0.301	0.301
2	14156	0.176	0.176
3	10227	0.127	0.125
4	7931	0.099	0.097
5	6359	0.079	0.079
6	5372	0.067	0.067
7	4476	0.056	0.058
8	4092	0.051	0.051
9	3650	0.045	0.046

$3x + 1$ Data: random 10,000 digit number, $2|3x + 1$

241,344 iterations, $\chi^2 = 11.4$ (5% 15.5).

Digit	Number	Observed	Benford
1	72924	0.302	0.301
2	42357	0.176	0.176
3	30201	0.125	0.125
4	23507	0.097	0.097
5	18928	0.078	0.079
6	16296	0.068	0.067
7	13702	0.057	0.058
8	12356	0.051	0.051
9	11073	0.046	0.046

$5x + 1$ Data: random 10,000 digit number, $2^k || 5x + 1$

27,004 iterations, $\chi^2 = 1.8$ (5% 15.5).

Digit	Number	Observed	Benford
1	8154	0.302	0.301
2	4770	0.177	0.176
3	3405	0.126	0.125
4	2634	0.098	0.097
5	2105	0.078	0.079
6	1787	0.066	0.067
7	1568	0.058	0.058
8	1357	0.050	0.051
9	1224	0.045	0.046

$5x+1$ Data: random 10,000 digit number, $2|5x+1$

241,344 iterations, $\chi^2 = 3 \cdot 10^{-4}$ (5% 15.5).

Digit	Number	Observed	Benford
1	72652	0.301	0.301
2	42499	0.176	0.176
3	30153	0.125	0.125
4	23388	0.097	0.097
5	19110	0.079	0.079
6	16159	0.067	0.067
7	13995	0.058	0.058
8	12345	0.051	0.051
9	11043	0.046	0.046

Conclusions

Conclusions and Future Investigations

- Different systems exhibit Benford behavior.
- Ingredients of proofs (logarithms, equidistribution).
- Applications to fraud detection / data integrity.
- Current studies include dependent random variables (partitions and amalgamations) and quantifying convergence to Benford.

References

Bibliography

- Benford, *The law of anomalous numbers*, Proceedings of the American Philosophical Society **78** (1938), 551–572.
- P. Diaconis, *The distribution of leading digits and uniform distribution mod 1*, Ann. Probab. **5** (1979), 72–81.
- T. Hill, *The first-digit phenomenon*, American Scientist **86** (1996), 358–363.
- T. Hill, *A statistical derivation of the significant-digit law*, Statistical Science **10** (1996), 354–363.
- W. Hurlimann, *Benford's Law from 1881 to 2006: a bibliography*, `http://arxiv.org/abs/math/0607168`.
- D. Jang, J. U. Kang, A. Kruckman, J. Kudo and S. J. Miller, *Chains of distributions, hierarchical Bayesian models and Benford's Law*, Journal of Algebra, Number Theory: Advances and Applications, volume 1, number 1 (March 2009), 37–60. `http://arxiv.org/abs/0805.4226`.

Bibliography

- A. Kontorovich and S. J. Miller, *Benford's Law, Values of L-functions and the* $3x+1$ *Problem*, Acta Arith. **120** (2005), 269–297. http://arxiv.org/abs/math/0412003.
- D. Knuth, *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*, Addison-Wesley, third edition, 1997.
- J. Lagarias and K. Soundararajan, *Benford's Law for the* $3x+1$ *Function*, J. London Math. Soc. (2) **74** (2006), no. 2, 289–303. http://arxiv.org/pdf/math/0608208.pdf.
- E. Ley, *On the peculiar distribution of the U.S. Stock Indices Digits*, The American Statistician **50** (1996), no. 4, 311–313.
- S. J. Miller and M. Nigrini, *The Modulo* 1 *Central Limit Theorem and Benford's Law for Products*, International Journal of Algebra **2** (2008), no. 3, 119–130. http://arxiv.org/pdf/math/0607686.pdf.
- S. J. Miller and M. Nigrini, *Order statistics and Benford's law*, International Journal of Mathematics and Mathematical Sciences, Volume 2008 (2008), Article ID 382948, 19 pages. doi:10.1155/2008/382948. http://arxiv.org/abs/math/0601344.

Bibliography

- S. Newcomb, *Note on the frequency of use of the different digits in natural numbers*, Amer. J. Math. **4** (1881), 39-40.
- M. Nigrini, *Digital Analysis and the Reduction of Auditor Litigation Risk*. Pages 69–81 in *Proceedings of the 1996 Deloitte & Touche / University of Kansas Symposium on Auditing Problems*, ed. M. Ettredge, University of Kansas, Lawrence, KS, 1996.
- M. Nigrini, *The Use of Benford's Law as an Aid in Analytical Procedures*, Auditing: A Journal of Practice & Theory, **16** (1997), no. 2, 52–67.
- M. Nigrini and S. J. Miller, *Benford's Law applied to hydrology data – results and relevance to other geophysical data*, Mathematical Geology **39** (2007), no. 5, 469–490.
- M. Nigrini and S. J. Miller, *Data diagnostics using second order tests of Benford's Law*, Auditing: A Journal of Practice and Theory **28** (2009), no. 2, 305–324. doi: 10.2308/aud.2009.28.2.305
- R. Pinkham, *On the Distribution of First Significant Digits*, The Annals of Mathematical Statistics **32**, no. 4 (1961), 1223-1230.
- R. A. Raimi, *The first digit problem*, Amer. Math. Monthly **83** (1976), no. 7, 521–538.

L-functions and Random Matrix Theory

Good L-functions ($\zeta(s)$, full level cusp form)

L-function is good if:

- Euler product:

$$L(s,f) = \sum_{n=1}^{\infty} \frac{a_f(n)}{n^s} = \prod_{p \text{ prime}} \prod_{j=1}^{d} \left(1 - \alpha_{f,j}(p) p^{-s}\right)^{-1}.$$

- meromorphic continuation to $\mathbb{C}$, of finite order, at most finitely many poles (all on the line $\mathrm{Re}(s) = 1$).
- Functional equation: $\omega \in \mathbb{R}$, $G(s)$ prod Γ-fns:

$$e^{i\omega} G(s) L(s,f) = e^{-i\omega} \overline{G(1-\overline{s}) L(1-\overline{s})}.$$

Good L-functions

- For some $\aleph > 0$, $c \in \mathbb{C}$, $x \geq 2$:

$$\sum_{p \leq x} \frac{|a_f(p)|^2}{p} = \aleph \log\log x + c + O\left(\frac{1}{\log x}\right).$$

- The $\alpha_{f,j}(p)$ are (Ramanujan-Petersson) tempered: $|\alpha_{f,j}(p)| \leq 1$.
- $N(\sigma, T) = \#\{\rho : L(\rho, f) = 0,\ \mathrm{Re}(\rho) \geq \sigma,\ \mathrm{Im}(\rho) \in [0, T]\}$. $\exists \beta > 0$

$$N(\sigma, T) = O\left(T^{1-\beta\left(\sigma-\frac{1}{2}\right)} \log T\right).$$

Log-Normal Law (Hejhal, Laurinčikas, Selberg)

Log-Normal Law

$$\frac{\mu\left(\{t \in [T, 2T] : \log\left|L\left(\sigma + it, f\right)\right| \in [a, b]\}\right)}{T} =$$

$$\frac{1}{\sqrt{\psi(\sigma, T)}} \int_a^b e^{-\pi u^2/\psi(\sigma,T)} du + \mathrm{Error}$$

$$\psi(\sigma, T) = \aleph \log\left[\min\left(\log T, \frac{1}{\sigma - \frac{1}{2}}\right)\right] + O(1)$$

$$\frac{1}{2} \le \sigma \le \frac{1}{2} + \frac{1}{\log^\delta T}, \quad \delta \in (0, 1).$$

Values of L-functions and Benford's Law

Theorem (Kontorovich and M–, 2005)

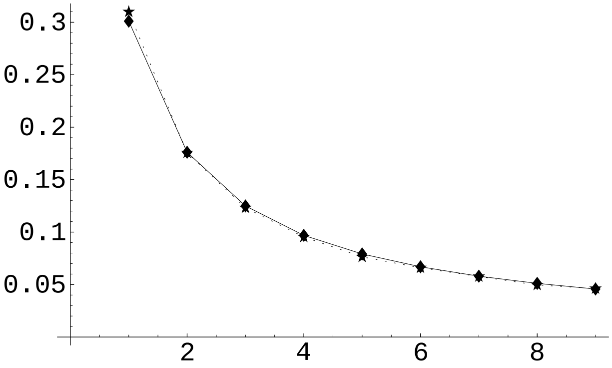
$L(s,f)$ a good L-function, as $T\to\infty$, $L(\sigma_T+it,f)$ is Benford.

Ingredients

- Approximate $\log L(\sigma_T+it,f)$ with $\sum_{n\le x}\frac{c(n)\Lambda(n)}{\log n}\frac{1}{n^{\sigma_T+it}}$.
- study moments $\int_T^{2T}|\cdot|$, $k\le\log^{1-\delta}T$.
- Montgomery-Vaughan: $\int_T^{2T}\sum a_n n^{-it}\overline{\sum b_m m^{-it}}dt = H\sum a_n\overline{b}_n + O(1)\sqrt{\sum n|a_n|^2\sum n|b_n|^2}$.

Riemann Zeta Function

$$\left|\zeta\left(\frac{1}{2}+i\frac{k}{4}\right)\right|,\ k\in\{0,1,\dots,65535\}.$$



Random Matrices: Preliminaries

- $N \times N$ unitary matrices U (Haar measure):

$$p_N(U) = \frac{1}{(2\pi)^N N!} \prod_{1 \le j < m \le N} \left| e^{\mathrm{i}\theta_j} - e^{\mathrm{i}\theta_m} \right|.$$

- characteristic polynomial:

$$Z(U,\theta) = \det(I - Ue^{-\mathrm{i}\theta}) = \prod \left(1 - e^{\mathrm{i}(\theta_n - \theta)}\right).$$

- $\rho_N(x)$ the probability density for $\log |Z(U,\theta)|$:

$$\widetilde{\rho}_N(x) \ = \ \sqrt{Q_2(N)}\, \rho_N(\sqrt{Q_2(N)}\, x),$$

variance $Q_2(N) \sim (\log N)/2$.

Random Matrices and Benford's Law

Theorem (Kontorovich and M–, 2005)

As $N \to \infty$, the distribution of digits of the absolute values of the characteristic polynomials of $N \times N$ unitary matrices (with respect to Haar measure) converges to the Benford probabilities.

- Key Ingredient: Keating-Snaith:

$$\widetilde{\rho}_N(x)dx \;=\; \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx + O\left((\log N)^{-3/2}dx\right).$$