Benford's law, or: Why the IRS cares about number theory!

Steven J Miller (Williams College)

sjm1@williams.edu
http://www.williams.edu/go/math/sjmiller/

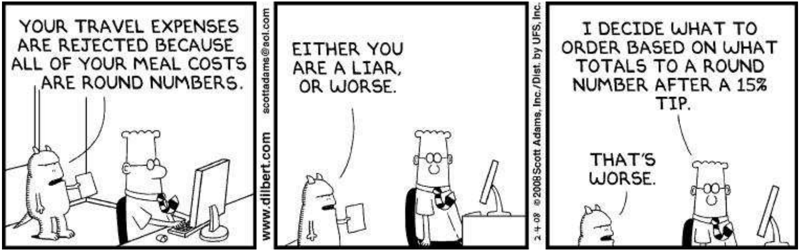
Bentley University, February 1, 2010

Summary

- Review 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.



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

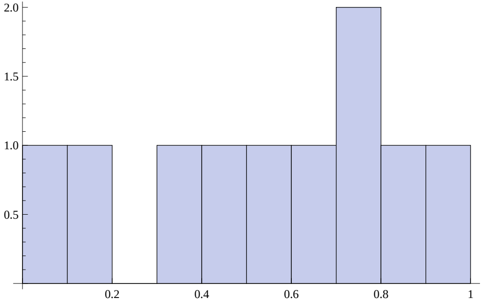
Equidistribution

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

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

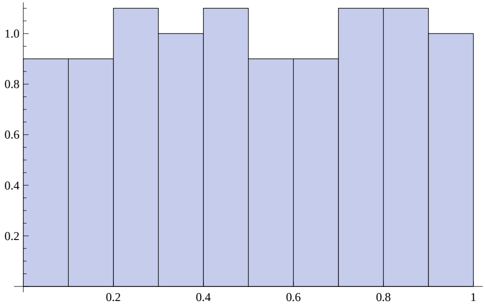
- Thm: $\beta \notin \mathbb{Q}$, $n\beta$ is equidistributed mod 1.
- Examples: $\log_{10} 2, \log_{10}\left(\frac{1+\sqrt{5}}{2}\right) \notin \mathbb{Q}$.
  Proof: if rational: $2 = 10^{p/q}$.
  Thus $2^q = 10^p$ or $2^{q-p} = 5^p$, impossible.

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



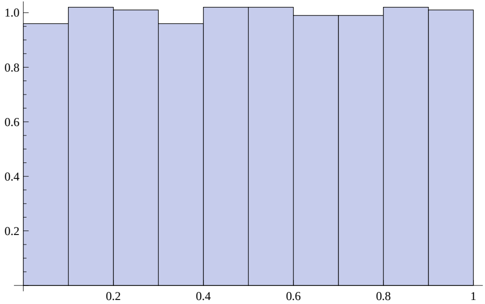
$n\sqrt{\pi}$ mod 1 for $n \leq 10$

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



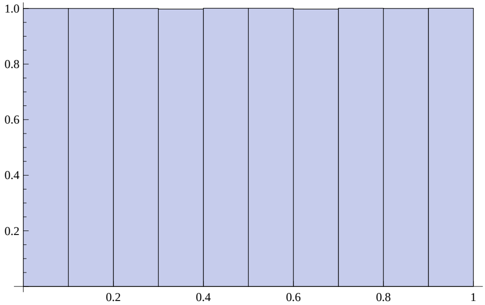
$n\sqrt{\pi} \bmod 1$ for $n \leq 100$

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



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

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



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

Logarithms and Benford's Law

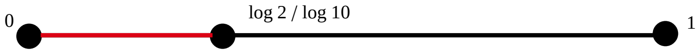
Fundamental Equivalence

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

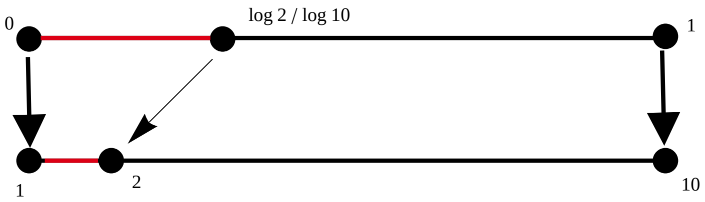
Data set $\{x_i\}$ is Benford base B if $\{y_i\}$ is equidistributed mod 1, where $y_i = \log_B x_i$.



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 $\log_{10} 2 \notin \mathbb{Q}$.
- Fibonacci numbers are Benford base 10.

  $a_{n+1} = a_n + a_{n-1}$.
  Guess $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ or $r^2 = r + 1$.
  Roots $r = (1 \pm \sqrt{5})/2$.
  General solution: $a_n = c_1 r_1^n + c_2 r_2^n$.
  Binet: $a_n = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2}\right)^n$.

- Most linear recurrence relations Benford:

Examples

- Fibonacci numbers are Benford base 10.

  $a_{n+1} = a_n + a_{n-1}$.
  Guess $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ or $r^2 = r + 1$.
  Roots $r = (1 \pm \sqrt{5})/2$.
  General solution: $a_n = c_1 r_1^n + c_2 r_2^n$.
  Binet: $a_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n$.

- Most linear recurrence relations Benford:

Examples

- Fibonacci numbers are Benford base 10.

  $a_{n+1} = a_n + a_{n-1}$.
  Guess $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ or $r^2 = r + 1$.
  Roots $r = (1 \pm \sqrt{5})/2$.
  General solution: $a_n = c_1 r_1^n + c_2 r_2^n$.
  Binet: $a_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n$.

- Most linear recurrence relations Benford:

Examples

- Fibonacci numbers are Benford base 10.

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

  Guess $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ or $r^2 = r + 1$.

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

  General solution: $a_n = c_1 r_1^n + c_2 r_2^n$.

  Binet: $a_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n$.

- Most linear recurrence relations Benford:

Examples

- Fibonacci numbers are Benford base 10.
  $a_{n+1} = a_n + a_{n-1}$.
  Guess $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ or $r^2 = r + 1$.
  Roots $r = (1 \pm \sqrt{5})/2$.
  General solution: $a_n = c_1 r_1^n + c_2 r_2^n$.
  Binet: $a_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n$.

- Most linear recurrence relations Benford:

Examples

- Fibonacci numbers are Benford base 10.

  $a_{n+1} = a_n + a_{n-1}$.
  Guess $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ or $r^2 = r + 1$.
  Roots $r = (1 \pm \sqrt{5})/2$.
  General solution: $a_n = c_1 r_1^n + c_2 r_2^n$.
  Binet: $a_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n$.

- Most linear recurrence relations Benford:

Examples

- Fibonacci numbers are Benford base 10.

  $a_{n+1} = a_n + a_{n-1}$.
  Guess $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ or $r^2 = r + 1$.
  Roots $r = (1 \pm \sqrt{5})/2$.
  General solution: $a_n = c_1 r_1^n + c_2 r_2^n$.
  Binet: $a_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n$.

- Most linear recurrence relations Benford:

Examples

- Fibonacci numbers are Benford base 10.

  $a_{n+1} = a_n + a_{n-1}$.
  Guess $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ or $r^2 = r + 1$.
  Roots $r = (1 \pm \sqrt{5})/2$.
  General solution: $a_n = c_1 r_1^n + c_2 r_2^n$.
  Binet: $a_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n$.

- Most linear recurrence relations Benford:

Examples

- Fibonacci numbers are Benford base 10.
  $a_{n+1} = a_n + a_{n-1}$.
  Guess $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ or $r^2 = r + 1$.
  Roots $r = (1 \pm \sqrt{5})/2$.
  General solution: $a_n = c_1 r_1^n + c_2 r_2^n$.
  Binet: $a_n = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2}\right)^n$.

- 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 $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ or $r^2 = r + 1$.
  Roots $r = (1 \pm \sqrt{5})/2$.
  General solution: $a_n = c_1 r_1^n + c_2 r_2^n$.
  Binet: $a_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n$.

- 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 $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ or $r^2 = r + 1$.
  Roots $r = (1 \pm \sqrt{5})/2$.
  General solution: $a_n = c_1 r_1^n + c_2 r_2^n$.
  Binet: $a_n = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2}\right)^n$.

- 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	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				

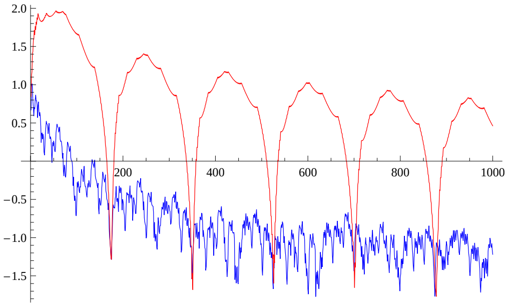
Logarithms and Benford's Law

χ^2 values for α^n, $1 \le n \le N$ (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

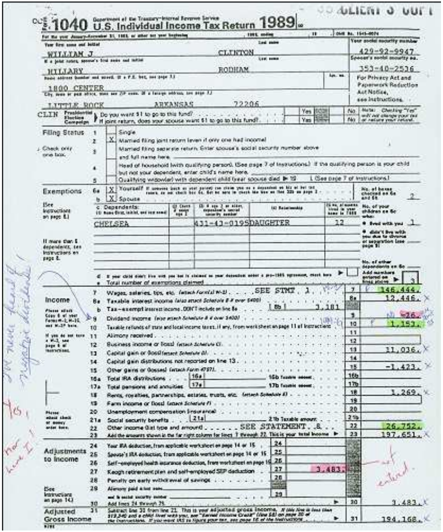
Logarithms and Benford's Law: Base 10

$\log(\chi^2)$ 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

Filing Status

Exemptions

CHELSEA 431-43-0195 DAUGHTER 12

Income

7 Wages, salaries, tips, etc. SEE STMT 1 146,444.

8a Taxable interest income 12,446.

9 Dividend income 26.

10 Taxable refunds of state and local income taxes 1,153.

13 Capital gain or (loss) 11,036.

15 Other gains or (losses) -1,423.

18 Rents, royalties, partnerships, estates, trusts, etc. 1,269.

22 Other income 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

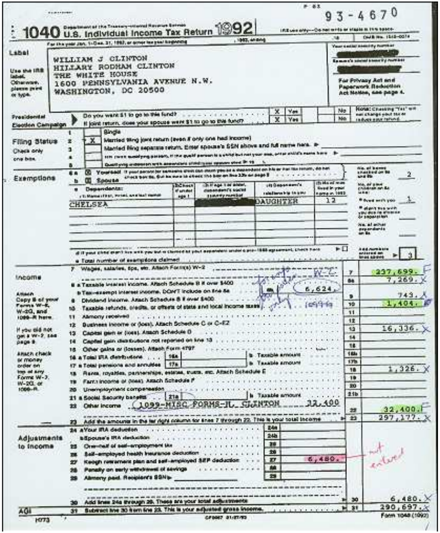
27 Keogh retirement plan and self-employed SEP deduction 3,483.

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

For the year Jan. 1–Dec. 31, 1992, or other tax year beginning , 1992, ending

Label — Use the IRS label. Otherwise, please print or type.

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 No
If joint return, does your spouse want $1 to go to this fund? X Yes No

Filing Status (Check only one box.)
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:

(1) Name (first, initial, and last name)	(2) Check if under age 1	(3) If age 1 or older, dependent's social security number	(4) Dependent's relationship to you	(5) No. of months lived in your home in 1992
CHELSEA		[redacted]	DAUGHTER	12

No. of boxes checked on 6a and 6b: 2
No. of your children on 6c who lived with you: 1
e Total number of exemptions claimed — Add numbers entered on lines above: 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 — 8b 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 — b Taxable amount
17a Total pensions and annuities — b Taxable amount
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 — b Taxable amount
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. Recipient's SSN
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.

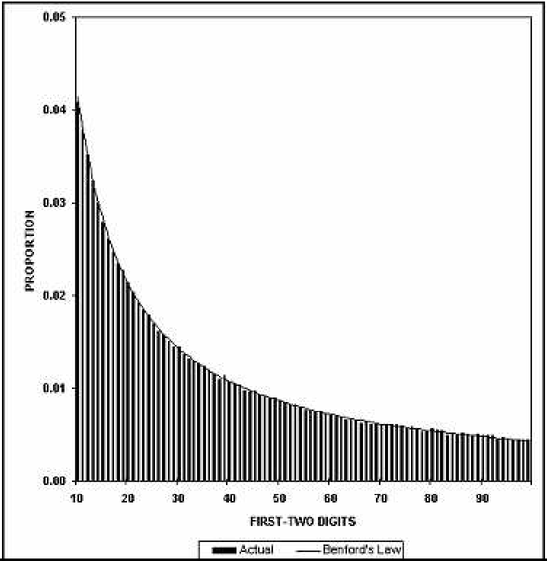
Form 1040 (1992)

Detecting Fraud

Bank Fraud

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

Data Integrity: Stream Flow Statistics: 130 years, 457,440 records



Election Fraud: Iran 2009

Numerous protests/complaints over Iran's 2009 elections.
Lot of analysis; data moderately suspicious:

- First and second leading digits;
- Last two digits (should almost be uniform);
- Last two digits differing by at least 2.

Warning: enough tests, even if nothing wrong will find a suspicious result (but when all tests are on the boundary...).

Benford Good 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 Formula: f nice:

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

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 ($h(|T|) \to \infty$):

- 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.

- 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

- Control translated errors:
  - ⋄ hardest step
  - ⋄ techniques problem specific

Sketch of the proof (continued)

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

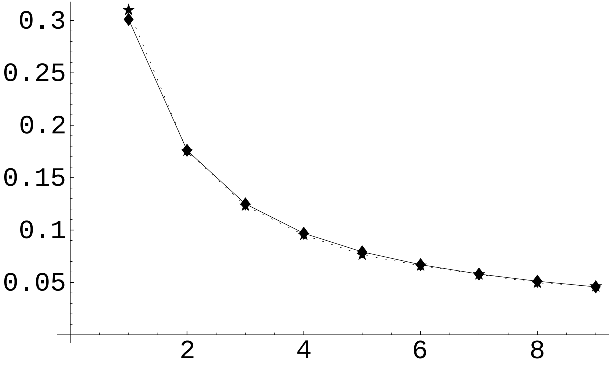
$$= \sum_{|\ell| \le Th(T)} \left[G_T(b+\ell) - G_T(a+\ell)\right] + o(1)$$

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

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

Riemann Zeta Function

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



Products of Random Variables

Preliminaries

- $X_1 \cdots X_n \Leftrightarrow Y_1 + \cdots + Y_n \bmod 1$, $Y_i = \log_B X_i$

- Density Y_i is g_i, density $Y_i + Y_j$ is

$$(g_i * g_j)(y) \;=\; \int_0^1 g_i(t) g_j(y-t) dt.$$

- $h_n = g_1 * \cdots * g_n$, $\widehat{g}(\xi) = \widehat{g}_1(\xi) \cdots \widehat{g}_n(\xi)$.

Modulo 1 Central Limit Theorem

Theorem (M– and Nigrini 2007)

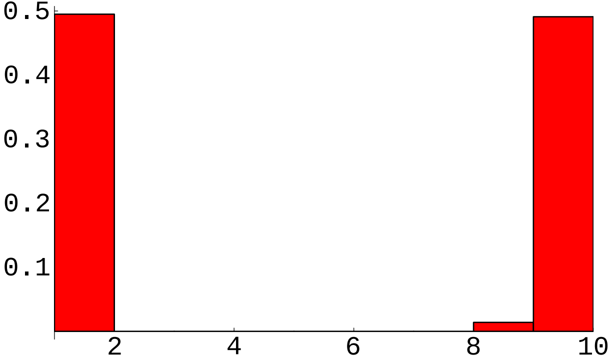
$\{Y_m\}$ independent continuous random variables on $[0,1)$ (not necc. i.i.d.), densities $\{g_m\}$. $Y_1 + \cdots + Y_M \bmod 1$ converges to the uniform distribution as $M \to \infty$ in $L^1([0,1])$ if and only if for all $n \neq 0$, $\lim_{M\to\infty} \widehat{g_1}(n) \cdots \widehat{g_M}(n) = 0$.

◇ Gives info on rate of convergence.

Generalizations

- Levy proved for i.i.d.r.v. just one year after Benford's paper.
- Generalized to other compact groups, with estimates on the rate of convergence.
  ◇ Stromberg: n-fold convolution of a regular probability measure on a compact Hausdorff group G converges to normalized Haar measure in weak-star topology iff support of the distribution not contained in a coset of a proper normal closed subgroup of G.

Distribution of digits (base 10) of 1000 products $X_1 \cdots X_{1000}$, where $g_{10,m} = \phi_{11^m}$.
$\phi_m(x) = m$ if $|x - 1/8| \leq 1/2m$ (0 otherwise).



Proof under stronger conditions

- Use standard CLT to show $Y_1 + \cdots + Y_M$ tends to a Gaussian.
- Use Poisson Summation to show the Gaussian tends to the uniform modulo 1.

Proof under stronger conditions

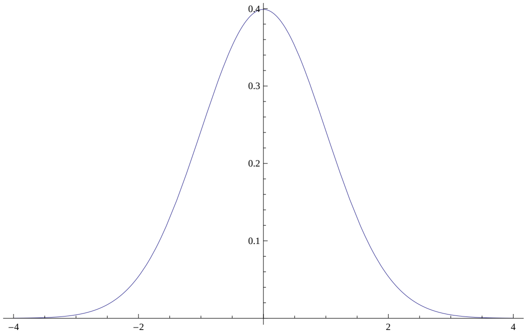


Figure: Plot of normal (mean 0, stdev 1).

Proof under stronger conditions

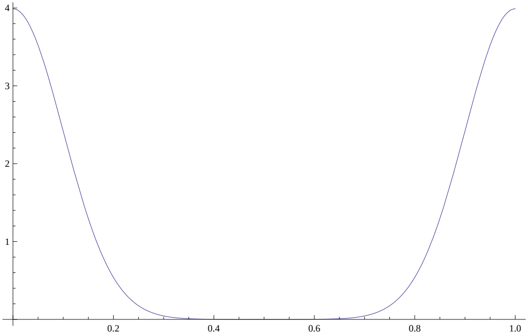


Figure: Plot of normal (mean 0, stdev .1) modulo 1.

Proof under stronger conditions

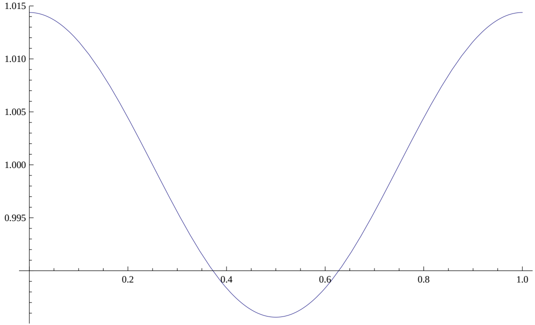


Figure: Plot of normal (mean 0, stdev .5) modulo 1.

Inputs

Poisson Summation Formula

f nice:

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

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

Lemma

$$\frac{2}{\sqrt{2\pi\sigma^2}} \int_{\sigma^{1+\delta}}^{\infty} e^{-x^2/2\sigma^2} dx \ll e^{-\sigma^{2\delta}/2}.$$

Proof Under Weaker Conditions

Lemma

As $N \to \infty$, $p_N(x) = \frac{e^{-\pi x^2/N}}{\sqrt{N}}$ becomes equidistributed modulo 1.

- $\int_{\substack{x=-\infty \\ x \bmod 1 \in [a,b]}}^{\infty} p_N(x)dx = \frac{1}{\sqrt{N}} \sum_{n\in\mathbb{Z}} \int_{x=a}^{b} e^{-\pi(x+n)^2/N} dx$.
- $e^{-\pi(x+n)^2/N} = e^{-\pi n^2/N} + O\left(\frac{\max(1,|n|)}{N} e^{-n^2/N}\right)$.
- Can restrict sum to $|n| \le N^{5/4}$.
- $\frac{1}{\sqrt{N}} \sum_{n\in\mathbb{Z}} e^{-\pi n^2/N} = \sum_{n\in\mathbb{Z}} e^{-\pi n^2 N}$.

Proof Under Weaker Conditions

$$\frac{1}{\sqrt{N}} \sum_{|n| \le N^{5/4}} \int_{x=a}^{b} e^{-\pi (x+n)^2/N} dx$$

$$= \frac{1}{\sqrt{N}} \sum_{|n| \le N^{5/4}} \int_{x=a}^{b} \left[e^{-\pi n^2/N} + O\left(\frac{\max(1,|n|)}{N} e^{-n^2/N} \right) \right] dx$$

$$= \frac{b-a}{\sqrt{N}} \sum_{|n| \le N^{5/4}} e^{-\pi n^2/N} + O\left(\frac{1}{N} \sum_{n=0}^{N^{5/4}} \frac{n+1}{\sqrt{N}} e^{-\pi (n/\sqrt{N})^2} \right)$$

$$= \frac{b-a}{\sqrt{N}} \sum_{|n| \le N^{5/4}} e^{-\pi n^2/N} + O\left(\frac{1}{N} \int_{w=0}^{N^{3/4}} (w+1) e^{-\pi w^2} \sqrt{N} dw \right)$$

$$= \frac{b-a}{\sqrt{N}} \sum_{|n| \le N^{5/4}} e^{-\pi n^2/N} + O\left(N^{-1/2} \right).$$

Proof Under Weaker Conditions

Extend sums to $n \in \mathbb{Z}$, apply Poisson Summation:

$$\frac{1}{\sqrt{N}} \sum_{n \in \mathbb{Z}} \int_{x=a}^{b} e^{-\pi(x+n)^2/N} dx \ \approx \ (b-a) \cdot \sum_{n \in \mathbb{Z}} e^{-\pi n^2 N}.$$

For $n = 0$ the right hand side is $b - a$.
For all other n, we trivially estimate the sum:

$$\sum_{n \neq 0} e^{-\pi n^2 N} \ \leq \ 2 \sum_{n \geq 1} e^{-\pi n N} \ \leq \ \frac{2e^{-\pi N}}{1 - e^{-\pi N}},$$

which is less than $4e^{-\pi N}$ for N sufficiently large.

Proof in General Case: Fourier input

- Fejér kernel:

$$F_N(x) = \sum_{n=-N}^{N} \left(1 - \frac{|n|}{N}\right) e^{2\pi inx}.$$

- Fejér series $T_N f(x)$ equals

$$(f * F_N)(x) = \sum_{n=-N}^{N} \left(1 - \frac{|n|}{N}\right) \widehat{f}(n) e^{2\pi inx}.$$

- Lebesgue's Theorem: $f \in L^1([0,1])$. As $N \to \infty$, $T_N f$ converges to f in $L^1([0,1])$.
- $T_N(f * g) = (T_N f) * g$: convolution assoc.

Proof of Modulo 1 CLT

- Density of sum is $h_\ell = g_1 * \cdots * g_\ell$.
- Suffices show $\forall\epsilon$: $\lim_{M\to\infty} \int_0^1 |h_M(x) - 1| dx < \epsilon$.
- Lebesgue's Theorem: N large,

$$||h_1 - T_N h_1||_1 = \int_0^1 |h_1(x) - T_N h_1(x)| dx < \frac{\epsilon}{2}.$$

- Claim: above holds for h_M for all M.

Proof of Modulo 1 CLT : Proof of Claim

$$T_N h_{M+1} = T_N(h_M * g_{M+1}) = (T_N h_M) * g_{M+1}$$

$$\begin{aligned}
||h_{M+1} - T_N h_{M+1}||_1 &= \int_0^1 |h_{M+1}(x) - T_N h_{M+1}(x)| dx \\
&= \int_0^1 |(h_M * g_{M+1})(x) - (T_N h_M) * g_{M+1}(x)| dx \\
&= \int_0^1 \left| \int_0^1 (h_M(y) - T_N h_M(y))\, g_{M+1}(x-y) \right| dydx \\
&\leq \int_0^1 \int_0^1 |h_M(y) - T_N h_M(y)| g_{M+1}(x-y) dxdy \\
&= \int_0^1 |h_M(y) - T_N h_M(y)| dy \cdot 1 \;<\; \frac{\epsilon}{2}.
\end{aligned}$$

Proof of Modulo 1 CLT

Show $\lim_{M\to\infty} ||h_M - 1||_1 = 0$.
Triangle inequality:

$$||h_M - 1||_1 \le ||h_M - T_N h_M||_1 + ||T_N h_M - 1||_1.$$

Choices of N and ϵ:

$$||h_M - T_N h_M||_1 < \epsilon/2.$$

Show $||T_N h_M - 1||_1 < \epsilon/2$.

Proof of Modulo 1 CLT

$$||T_N h_M - 1||_1 = \int_0^1 \left| \sum_{\substack{n=-N \\ n\neq 0}}^{N} \left(1 - \frac{|n|}{N}\right) \widehat{h_M}(n) e^{2\pi i n x} \right| dx$$

$$\leq \sum_{\substack{n=-N \\ n\neq 0}}^{N} \left(1 - \frac{|n|}{N}\right) |\widehat{h_M}(n)|$$

$\widehat{h_M}(n) = \widehat{g_1}(n) \cdots \widehat{g_M}(n) \longrightarrow_{M\to\infty} 0$.
For fixed N and ϵ, choose M large so that $|\widehat{h_M}(n)| < \epsilon/4N$ whenever $n \neq 0$ and $|n| \leq N$.

Products
and
Chains of Random Variables

Key Ingredients

- Mellin transform and Fourier transform related by logarithmic change of variable.

- Poisson summation from collapsing to modulo 1 random variables.

Preliminaries

- $\Xi_1, \dots, \Xi_n$ nice independent r.v.'s on $[0, \infty)$.
- Density $\Xi_1 \cdot \Xi_2$:

$$\int_0^\infty f_2\left(\frac{x}{t}\right) f_1(t) \frac{dt}{t}$$

$\diamond$ Proof: $\mathrm{Prob}(\Xi_1 \cdot \Xi_2 \in [0, x])$:

$$\int_{t=0}^\infty \mathrm{Prob}\left(\Xi_2 \in \left[0, \frac{x}{t}\right]\right) f_1(t) dt$$

$$= \int_{t=0}^\infty F_2\left(\frac{x}{t}\right) f_1(t) dt,$$

differentiate.

Mellin Transform

$$(\mathcal{M}f)(s) = \int_0^\infty f(x) x^s \frac{dx}{x}$$

$$(\mathcal{M}^{-1}g)(x) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} g(s) x^{-s} ds$$

$$g(s) = (\mathcal{M}f)(s), f(x) = (\mathcal{M}^{-1}g)(x).$$

$$(f_1 \star f_2)(x) = \int_0^\infty f_2\left(\frac{x}{t}\right) f_1(t) \frac{dt}{t}$$

$$(\mathcal{M}(f_1 \star f_2))(s) = (\mathcal{M}f_1)(s) \cdot (\mathcal{M}f_2)(s).$$

Mellin Transform Formulation: Products Random Variables

Theorem

X_i's independent, densities f_i. $\Xi_n = X_1 \cdots X_n$,

$$\begin{aligned} h_n(x_n) &= (f_1 \star \cdots \star f_n)(x_n) \\ (\mathcal{M}h_n)(s) &= \prod_{m=1}^{n} (\mathcal{M}f_m)(s). \end{aligned}$$

As $n \to \infty$, Ξ_n becomes Benford: $Y_n = \log_B \Xi_n$,
$|\mathrm{Prob}(Y_n \bmod 1 \in [a,b]) - (b-a)| \leq$

$$(b-a) \cdot \sum_{\ell \neq 0, \ell=-\infty}^{\infty} \prod_{m=1}^{n} (\mathcal{M}f_i)\left(1 - \frac{2\pi i \ell}{\log B}\right).$$

Proof of Kossovsky's Chain Conjecture for certain densities

Conditions

- $\{\mathcal{D}_i(\theta)\}_{i\in I}$: one-parameter distributions, densities $f_{\mathcal{D}_i(\theta)}$ on $[0,\infty)$.
- $p:\mathbb{N}\to I$, $X_1 \sim \mathcal{D}_{p(1)}(1)$, $X_m \sim \mathcal{D}_{p(m)}(X_{m-1})$.
- $m \geq 2$,

$$f_m(x_m) \;=\; \int_0^\infty f_{\mathcal{D}_{p(m)}(1)}\left(\frac{x_m}{x_{m-1}}\right) f_{m-1}(x_{m-1}) \frac{dx_{m-1}}{x_{m-1}}$$

-

$$\lim_{n\to\infty} \sum_{\substack{\ell=-\infty \\ \ell\neq 0}}^{\infty} \prod_{m=1}^{n} (\mathcal{M}f_{\mathcal{D}_{p(m)}(1)})\left(1-\frac{2\pi i\ell}{\log B}\right) \;=\; 0$$

Chains of Random Variables

Return to street problem: chain of uniforms.

Let $\mathcal{D}_{\rm unif}(\theta)$ be the density of a uniform random variable on $[0, \theta]$.

Let $X_1 \sim \mathcal{D}_{\rm unif}(1)$ and $X_{n+1} \sim \mathcal{D}_{\rm unif}(X_n)$.

Proof of Kossovsky's Chain Conjecture for certain densities

Theorem (JKKKM)

- *If conditions hold, as $n \to \infty$ the distribution of leading digits of X_n tends to Benford's law.*
- *The error is a nice function of the Mellin transforms: if $Y_n = \log_B X_n$, then*

$$\left|\text{Prob}(Y_n \bmod 1 \in [a, b]) - (b + a)\right| \le$$
$$\left|(b-a) \cdot \sum_{\substack{\ell=-\infty \\ \ell \neq 0}}^{\infty} \prod_{m=1}^{n} \left(\mathcal{M}f_{\mathcal{D}_{p(m)}}(1)\right) \left(1 - \frac{2\pi i \ell}{\log B}\right)\right|$$

Example: All $X_i \sim \mathrm{Exp}(1)$

- $X_i \sim \mathrm{Exp}(1)$, $Y_n = \log_B \Xi_n$.
- Needed ingredients:
  - ⋄ $\int_0^\infty \exp(-x) x^{s-1} dx = \Gamma(s)$.
  - ⋄ $|\Gamma(1+ix)| = \sqrt{\pi x / \sinh(\pi x)}, x \in \mathbb{R}$.
- $|P_n(s) - \log_{10}(s)| \leq$

$$\log_B s \sum_{\ell=1}^{\infty} \left(\frac{2\pi^2 \ell / \log B}{\sinh(2\pi^2 \ell / \log B)} \right)^{n/2}.$$

Example: All $X_i \sim \mathrm{Exp}(1)$

Bounds on the error

- $|P_n(s) - \log_{10} s| \le$
  - $\diamond\ 3.3 \cdot 10^{-3} \log_B s$ if $n = 2$,
  - $\diamond\ 1.9 \cdot 10^{-4} \log_B s$ if $n = 3$,
  - $\diamond\ 1.1 \cdot 10^{-5} \log_B s$ if $n = 5$, and
  - $\diamond\ 3.6 \cdot 10^{-13} \log_B s$ if $n = 10$.
- Error at most

$$\log_{10} s \sum_{\ell=1}^{\infty} \left(\frac{17.148\ell}{\exp(8.5726\ell)} \right)^{n/2} \le .057^n \log_{10} s$$

Conclusions

Conclusions and Future Investigations

- See many different systems exhibit Benford behavior.
- Ingredients of proofs (logarithms, equidistribution).
- Applications to fraud detection / data integrity.
- Future work:
  - ◇ Study digits of other systems.
  - ◇ Develop more sophisticated tests for fraud

References

- A. K. Adhikari, *Some results on the distribution of the most significant digit*, Sankhyā: The Indian Journal of Statistics, Series B **31** (1969), 413–420.
- A. K. Adhikari and B. P. Sarkar, *Distribution of most significant digit in certain functions whose arguments are random variables*, Sankhyā: The Indian Journal of Statistics, Series B **30** (1968), 47–58.
- R. N. Bhattacharya, *Speed of convergence of the n-fold convolution of a probability measure ona compact group*, Z. Wahrscheinlichkeitstheorie verw. Geb. **25** (1972), 1–10.
- F. Benford, *The law of anomalous numbers*, Proceedings of the American Philosophical Society **78** (1938), 551–572.
- A. Berger, Leonid A. Bunimovich and T. Hill, *One-dimensional dynamical systems and Benford's Law*, Trans. Amer. Math. Soc. **357** (2005), no. 1, 197–219.

- A. Berger and T. Hill, *Newton's method obeys Benford's law*, The Amer. Math. Monthly **114** (2007), no. 7, 588-601.
- J. Boyle, *An application of Fourier series to the most significant digit problem* Amer. Math. Monthly **101** (1994), 879–886.
- J. Brown and R. Duncan, *Modulo one uniform distribution of the sequence of logarithms of certain recursive sequences*, Fibonacci Quarterly **8** (1970) 482–486.
- P. Diaconis, *The distribution of leading digits and uniform distribution mod 1*, Ann. Probab. **5** (1979), 72–81.
- W. Feller, *An Introduction to Probability Theory and its Applications, Vol. II*, second edition, John Wiley & Sons, Inc., 1971.

- R. W. Hamming, *On the distribution of numbers*, Bell Syst. Tech. J. **49** (1970), 1609-1625.
- 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.
- P. J. Holewijn, *On the uniform distribuiton of sequences of random variables*, Z. Wahrscheinlichkeitstheorie verw. Geb. **14** (1969), 89–92.
- 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.

- E. Janvresse and T. de la Rue, *From uniform distribution to Benford's law*, Journal of Applied Probability **41** (2004) no. 4, 1203–1210.
- A. Kontorovich and S. J. Miller, *Benford's Law, Values of L-functions and the* $3x+1$ *Problem*, Acta Arith. **120** (2005), 269–297.
- 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.
- S. Lang, *Undergraduate Analysis*, 2nd edition, Springer-Verlag, New York, 1997.

- P. Levy, *L'addition des variables aléatoires définies sur une circonférence*, Bull. de la S. M. F. **67** (1939), 1–41.
- E. Ley, *On the peculiar distribution of the U.S. Stock Indices Digits*, The American Statistician **50** (1996), no. 4, 311–313.
- R. M. Loynes, *Some results in the probabilistic theory of asympototic uniform distributions modulo* 1, Z. Wahrscheinlichkeitstheorie verw. Geb. **26** (1973), 33–41.
- S. J. Miller, *When the Cramér-Rao Inequality provides no information*, Communications in Information and Systems **7** (2007), no. 3, 265–272.
- 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.
- 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.

- S. J. Miller and R. Takloo-Bighash, *An Invitation to Modern Number Theory*, Princeton University Press, Princeton, NJ, 2006.
- 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.
- 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.
- H. Robbins, *On the equidistribution of sums of independent random variables*, Proc. Amer. Math. Soc. **4** (1953), 786–799.
- H. Sakamoto, *On the distributions of the product and the quotient of the independent and uniformly distributed random variables*, Tôhoku Math. J. **49** (1943), 243–260.
- P. Schatte, *On sums modulo 2π of independent random variables*, Math. Nachr. **110** (1983), 243–261.

- P. Schatte, *On the asymptotic uniform distribution of sums reduced mod 1*, Math. Nachr. **115** (1984), 275–281.
- P. Schatte, *On the asymptotic logarithmic distribution of the floating-point mantissas of sums*, Math. Nachr. **127** (1986), 7–20.
- E. Stein and R. Shakarchi, *Fourier Analysis: An Introduction*, Princeton University Press, 2003.
- M. D. Springer and W. E. Thompson, *The distribution of products of independent random variables*, SIAM J. Appl. Math. **14** (1966) 511–526.
- K. Stromberg, *Probabilities on a compact group*, Trans. Amer. Math. Soc. **94** (1960), 295–309.
- P. R. Turner, *The distribution of leading significant digits*, IMA J. Numer. Anal. **2** (1982), no. 4, 407–412.

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

$3x + 1$ Problem

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

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

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

Structure Theorem: Sinai, Kontorovich-Sinai

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

$(k_1, \dots, k_m)$: 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, $1/2$):

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.

Sketch of the proof

- Failed Proof: lattices, bad errors.
- CLT: $(S_m - 2m)/\sqrt{2m} \to N(0,1)$:

$$\mathbb{P}\left(S_m - 2m = k\right) = \frac{\eta(k/\sqrt{m})}{\sqrt{m}} + O\left(\frac{1}{g(m)\sqrt{m}}\right).$$

- Quantified Equidistribution: $I_\ell = \{\ell M, \dots, (\ell+1)M - 1\}$, $M = m^c$, $c < 1/2$
  $k_1, k_2 \in I_\ell$: $\left|\eta\left(\frac{k_1}{\sqrt{m}}\right) - \eta\left(\frac{k_2}{\sqrt{m}}\right)\right|$ small
  $C = \log_B 2$ of irrationality type $\kappa < \infty$:

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

Sketch of the proof: 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.

Sketch of the proof: 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 $\kappa < 1.2 \cdot 10^{602}$:

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

Sketch of the proof : Quantified Equidistribution

Theorem (Erdös-Turan)

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

There is a C such that for all m:

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

Sketch of the proof : 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 $m = \lfloor N^{1/\kappa} \rfloor$.

$3x+1$ 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