Why the IRS cares about the Riemann Zeta Function and Number Theory (and why you should too!)

Steven J. Miller
sjm1@williams.edu,
Steven.Miller.MC.96@aya.yale.edu

http://web.williams.edu/Mathematics/
sjmiller/public_html/

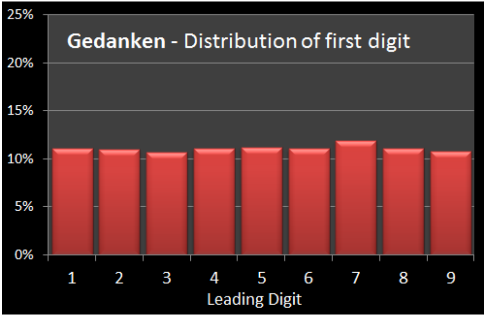
Stresa, Italy, July 11, 1, 2019

Introduction

- A. Berger and T. P. Hill, *An Introduction to Benford's Law*, Princeton University Press, Princeton, 2015. See also http://www.benfordonline.net/.
- A. E. Kossovsky, *Benford's Law: Theory, the General Law of Relative Quantities, and Forensic Fraud Detection Applications*, WSPC, 2014.
- S. J. Miller (editor), *Theory and Applications of Benford's Law*, Princeton University Press, 2015.
- M. Nigrini, *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*, 1st Edition, Wiley, 2014.

Interesting Question

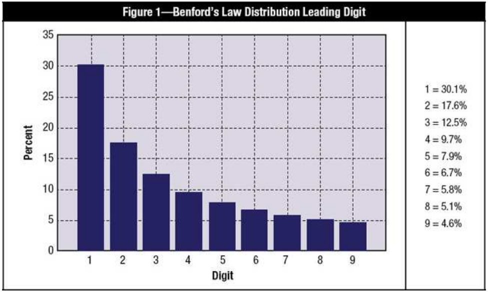
Motivating Question: For a nice data set, such as the Fibonacci numbers, stock prices, street addresses of college employees and students, ..., what percent of the leading digits are 1?



Natural guess: 10% (but immediately correct to 11%!).

Interesting Question

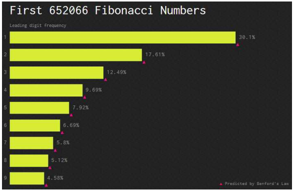
Motivating Question: For a nice data set, such as the Fibonacci numbers, stock prices, street addresses of college employees and students, ..., what percent of the leading digits are 1?



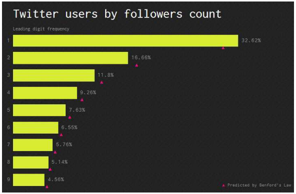
Answer: Benford's law!

Examples with First Digit Bias

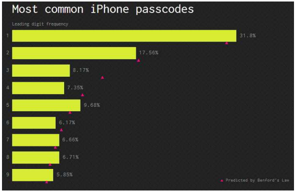
Fibonacci numbers



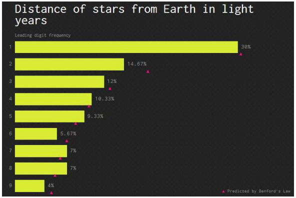
Twitter users by # followers



Most common iPhone passcodes



Distance of stars from Earth



Summary

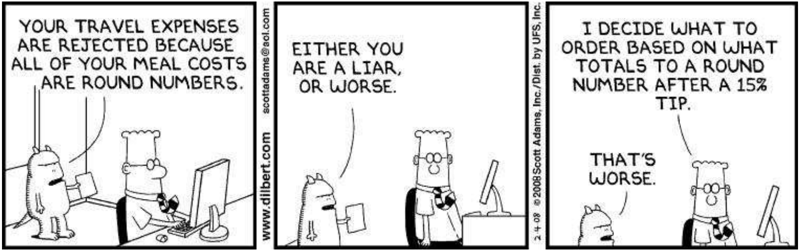
- Explain Benford's Law.
- Discuss examples and applications.
- Sketch proofs.
- Describe open problems.

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.



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

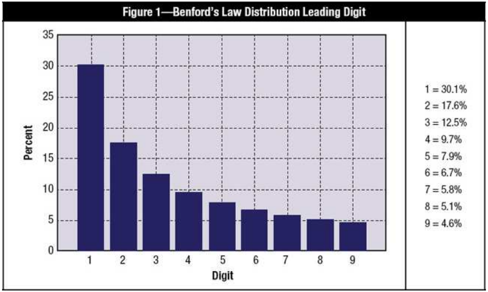
- Analyzing round-off errors.
- Determining the optimal way to store numbers.
- Detecting tax and image fraud, and data integrity.

General Theory

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 (probabilities)

Background Material

- Modulo: $a = b \mod c$ if $a - b$ is an integer times c; thus $17 = 5 \mod 12$, and $4.5 = .5 \mod 1$.

Background Material

- Modulo: $a = b \mod c$ if $a - b$ is an integer times c; thus $17 = 5 \mod 12$, and $4.5 = .5 \mod 1$.

- Significand: $x = S_{10}(x) \cdot 10^k$, k integer, $1 \leq S_{10}(x) < 10$.

Background Material

- Modulo: $a = b \mod c$ if $a - b$ is an integer times c; thus $17 = 5 \mod 12$, and $4.5 = .5 \mod 1$.
- Significand: $x = S_{10}(x) \cdot 10^k$, k integer, $1 \leq S_{10}(x) < 10$.
- $S_{10}(x) = S_{10}(\widetilde{x})$ if and only if x and $\widetilde{x}$ have the same leading digits. Note $\log_{10} x = \log_{10} S_{10}(x) + k$.

Background Material

- Modulo: $a = b \mod c$ if $a - b$ is an integer times c; thus $17 = 5 \mod 12$, and $4.5 = .5 \mod 1$.

- Significand: $x = S_{10}(x) \cdot 10^k$, k integer, $1 \leq S_{10}(x) < 10$.

- $S_{10}(x) = S_{10}(\widetilde{x})$ if and only if x and $\widetilde{x}$ have the same leading digits. Note $\log_{10} x = \log_{10} S_{10}(x) + k$.

- Key observation: $\log_{10}(x) = \log_{10}(\widetilde{x}) \mod 1$ if and only if x and $\widetilde{x}$ have the same leading digits.

Thus often study $y = \log_{10} x \mod 1$.
Advanced: $e^{2\pi i u} = e^{2\pi i (u \mod 1)}$.

Equidistribution and Benford's Law

Equidistribution

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

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

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 $b-a$:

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

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

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

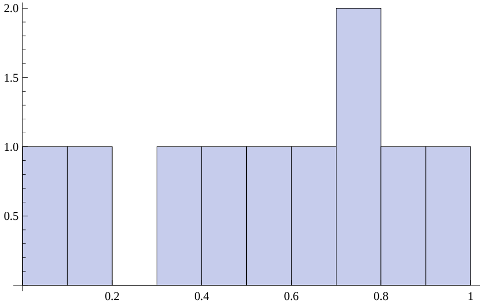
Equidistribution

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

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

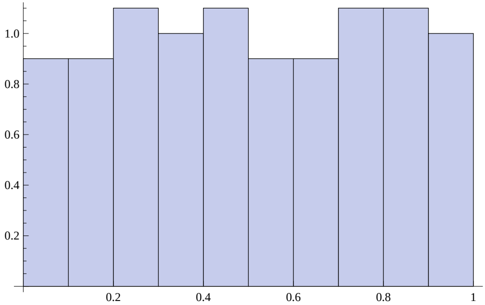
- Thm: $\beta \notin \mathbb{Q}$, $n\beta$ is equidistributed mod 1.
- Examples: $\log_{10} 2, \log_{10}\left(\frac{1+\sqrt{5}}{2}\right) \notin \mathbb{Q}$.
  Proof: if rational: $2 = 10^{p/q}$.
  Thus $2^q = 10^p$ or $2^{q-p} = 5^p$, impossible.

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



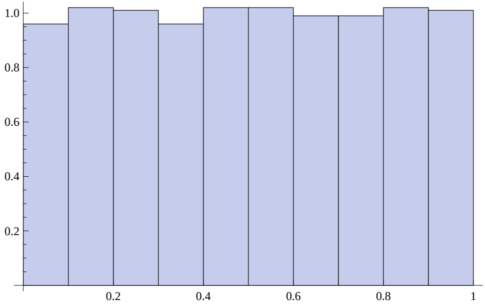
$n\sqrt{\pi} \bmod 1$ for $n \leq 10$

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



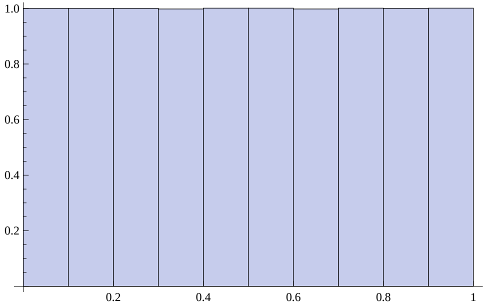
$n\sqrt{\pi} \bmod 1$ for $n \leq 100$

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



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

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



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

Logarithms and Benford's Law

Fundamental Equivalence

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

Logarithms and Benford's Law

Fundamental Equivalence

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

$$x = S_{10}(x) \cdot 10^k \text{ then}$$

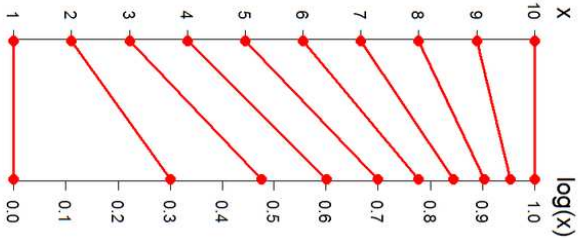
$$\log_{10} x = \log_{10} S_{10}(x) + k = \log_{10} S_{10} x \mod 1.$$

Logarithms and Benford's Law

Fundamental Equivalence

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

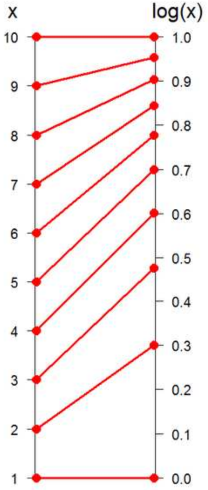
$$x = S_{10}(x) \cdot 10^k \text{ then}$$
$$\log_{10} x = \log_{10} S_{10}(x) + k = \log_{10} S_{10}x \mod 1.$$



Logarithms and Benford's Law

$$
\begin{aligned}
&\text{Prob}(\text{leading digit } d) \\
&= \log_{10}(d+1) - \log_{10}(d) \\
&= \log_{10}\left(\frac{d+1}{d}\right) \\
&= \log_{10}\left(1 + \frac{1}{d}\right).
\end{aligned}
$$

Have Benford's law $\leftrightarrow$ mantissa of logarithms of data are uniformly distributed



The Power of the Right Perspective

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	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	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

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				
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

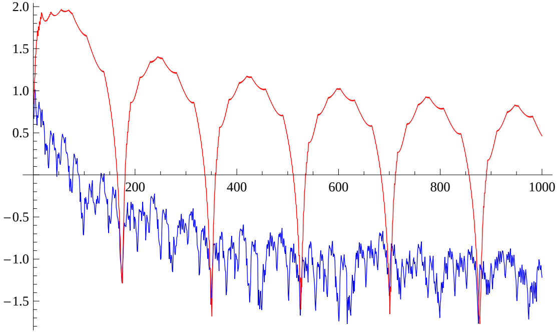
Logarithms and Benford's Law

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

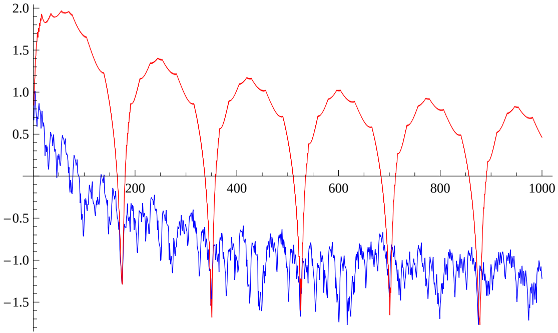
Logarithms and Benford's Law: Base 10 (5%: $\log(\chi^2) \approx 2.74$)

$\log(\chi^2)$ vs N for π^n (red) and e^n (blue), $n \in \{1, \dots, N\}$.



Logarithms and Benford's Law: Base 10 (5%: $\log(\chi^2) \approx 2.74$)

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



New Result: Linear Recurrence Relations of Degree 2

- $a_{n+1} = f(n)a_n + g(n)a_{n-1}$ with non-constant coefficients $f(n)$ and $g(n)$.
- Explore conditions on f and g such that the sequence generated obeys Benford's Law for all initial values.
- First solve the closed form of the sequence (a_n), then analyze its main term.

- Main idea: reduce the degree of recurrence.
- $a_{n+1} = (\lambda(n) + \mu(n))a_n - \mu(n)\lambda(n-1)a_{n-1}$, and compare the coefficients:

$$f(n) = \lambda(n) + \mu(n)$$
$$g(n) = -\lambda(n-1)\mu(n).$$

- We show that for any given pair of f and g, such λ and μ always exist.

Linear Recurrence Relations of Degree 2

- Recurrence relations of degree 1:

$$\begin{aligned} a_{n+1} &= \lambda(n) a_n + b_n \\ b_n &= \mu(n) b_{n-1}. \end{aligned}$$

- $a_{n+1} = r(n)\left(1 + \sum_{k=3}^{n} \prod_{i=k}^{n} \frac{\lambda(i)}{\mu(i)} + \frac{a_2}{b_1} \prod_{i=2}^{n} \frac{\lambda(i)}{\mu(i)}\right),$

  where $r(n) := b_1 \prod_{i=2}^{n} \mu(i)$.

- Find conditions on μ, λ such that main term dominates; Benford if $\prod \mu(i)$ is.

Examples when f and g are functions

- If $\mu(k) = k$, then $r(n) = n!$.
- If $\mu(k) = k^\alpha$ where $\alpha \in \mathbb{R}$, then $r(n) = (n!)^\alpha$.
- If $\mu(k) = \exp(\alpha h(k))$ where α is irrational and $h(k)$ is a monic polynomial, then $\log r(n) = \alpha \sum_{k=1}^{n} h(k)$.

Lemma

The sequence $\{\alpha p(n)\}$ is equidistributed mod 1 if $\alpha \notin \mathbb{Q}$ and $p(n)$ a monic polynomial.

Examples when f and g are random variables

- Take $\mu(n) \sim h(n)U_n$ where the U_n's are independent uniform distributions on $[0,1]$, and $h(n)$ is a deterministic function in n such that $\prod_{i=1}^{n} h(i)$ is Benford.

  Then $r(n) = \prod_{i=1}^{n} h(i) \prod_{i=1}^{n} U_i$ is Benford.

- Take $\mu(n) \sim \exp(U_n)$ where the U_n's are i.i.d. random variables. Then take logarithm and sum up $\log(\mu(n))$. Apply Central Limit Theorem and get a Gaussian distribution

Linear Recurrences of Higher Degree

- Use recurrence relation of degree 3 as an example. Similar main idea: reduce the degree.

- Define the sequence $\{a_n\}_{n=1}^{\infty}$ by $a_{n+1} \ = \ f_1(n)a_n + f_2(n)a_{n-1} + f_3(n)a_{n-2}$.

- Define an auxiliary sequence $(b_n)_{n=1}^{\infty}$ by $b_n \ = \ a_{n+1} - \lambda(n)a_n$. Then (b_n) is degree 2.

Why Benford's Law?

Streets

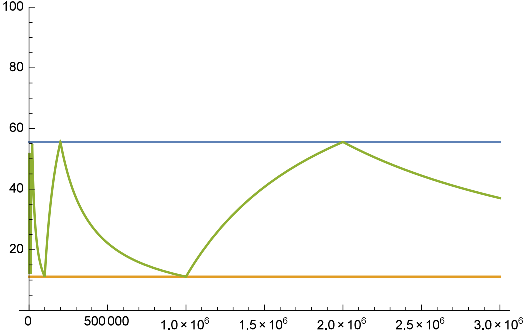
Not all data sets satisfy Benford's Law.

- Long street $[1, L]$: $L = 199$ versus $L = 999$.
- Oscillates b/w $1/9$ and $5/9$ with first digit 1.

Streets

Not all data sets satisfy Benford's Law.

- Long street $[1, L]$: $L = 199$ versus $L = 999$.
- Oscillates b/w $1/9$ and $5/9$ with first digit 1.

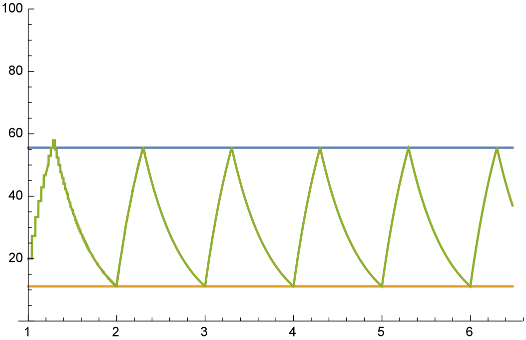


Probability first digit 1 versus street length L.

Streets

Not all data sets satisfy Benford's Law.

- Long street $[1, L]$: $L = 199$ versus $L = 999$.
- Oscillates b/w $1/9$ and $5/9$ with first digit 1.

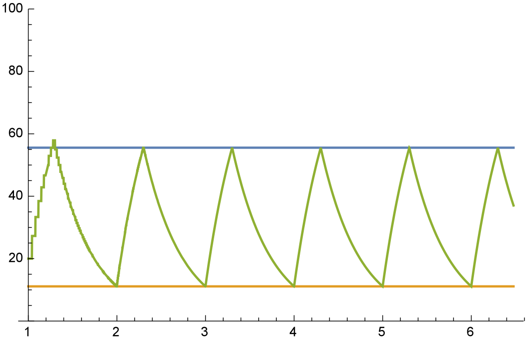


Probability first digit 1 versus log(street length L).

Streets

Not all data sets satisfy Benford's Law.

- Long street $[1, L]$: $L = 199$ versus $L = 999$.
- Oscillates b/w $1/9$ and $5/9$ with first digit 1.



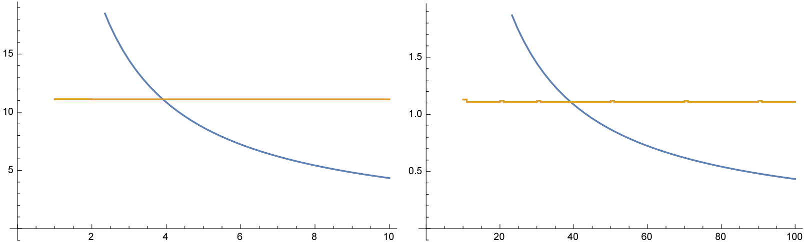
Probability first digit 1 versus log(street length L).

What if we have many streets of different lengths?

Amalgamating Streets

All houses: 1000 Streets,

each from 1 to 10000.

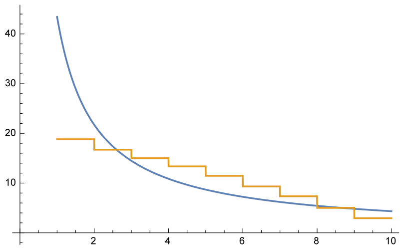


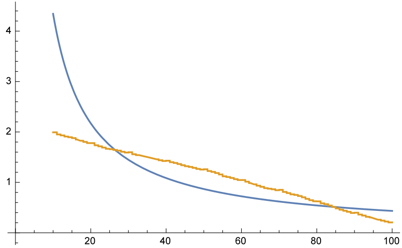
First digit and first two digits vs Benford.

Amalgamating Streets

All houses: 1000 Streets,

each from 1 to rand(10000).



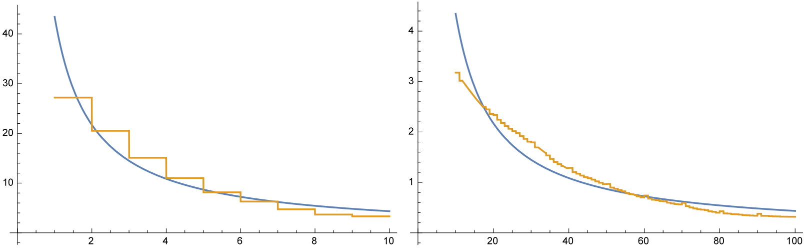


First digit and first two digits vs Benford.

Amalgamating Streets

All houses: 1000 Streets,

each 1 to rand(rand(10000)).



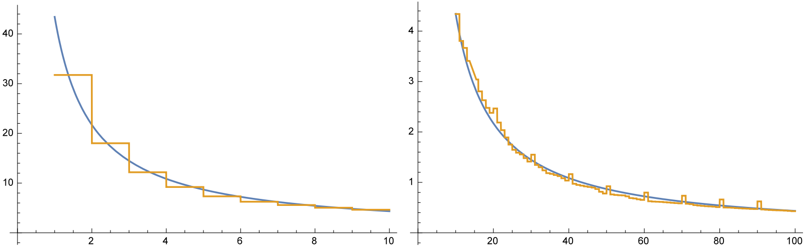
First digit and first two digits vs Benford.

Conclusion: More processes, closer to Benford.

Amalgamating Streets

All houses: 1000 Streets,

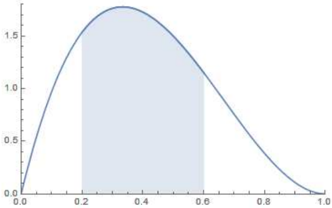
each 1 to rand(rand(rand(10000))).



First digit and first two digits vs Benford.

Conclusion: More processes, closer to Benford.

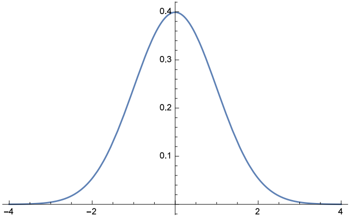
Probability Review



- Let X be random variable with density $p(x)$:
  - ◇ $p(x) \geq 0$; $\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$;
  - ◇ $\mathrm{Prob}\,(a \leq X \leq b) = \int_a^b p(x)\mathrm{d}x$.
- Mean $\mu = \int_{-\infty}^{\infty} x p(x)\mathrm{d}x$.
- Variance $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\mathrm{d}x$.
- Independence: knowledge of one random variable gives no knowledge of the other.

Central Limit Theorem

$$\text{Normal } N(\mu, \sigma^2): \ p(x) \ = \ e^{-(x-\mu)^2/2\sigma^2}/\sqrt{2\pi\sigma^2}.$$



Theorem

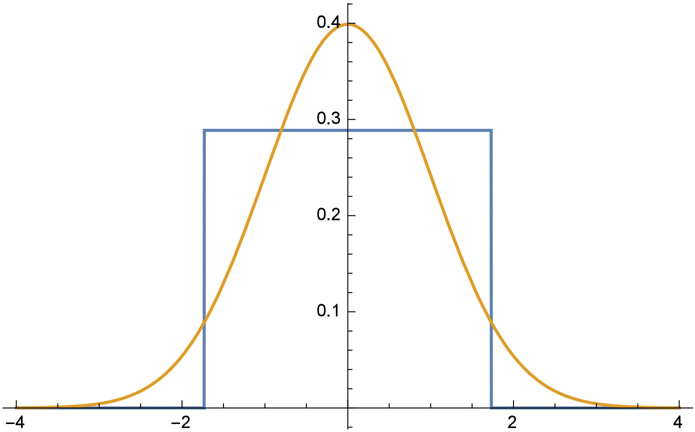
If $X_1, X_2, \dots$ independent, identically distributed random variables (mean μ, variance σ^2, finite moments) then

$$S_N \ := \ \frac{X_1 + \cdots + X_N - N\mu}{\sigma\sqrt{N}} \ \text{converges to} \ N(0,1).$$

Central Limit Theorem: Sums of Uniform Random Variables

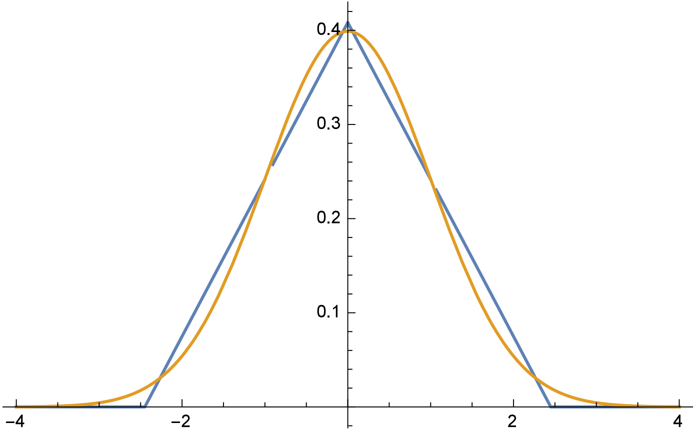
$X_i \sim \mathrm{Unif}(-1/2, 1/2)$ **(adjusted to mean 0, variance 1)**

$Y_1 = X_1/\sigma_{X_1}$ vs $N(0,1)$.



Central Limit Theorem: Sums of Uniform Random Variables
$X_i \sim \mathrm{Unif}(-1/2, 1/2)$ (adjusted to mean 0, variance 1)

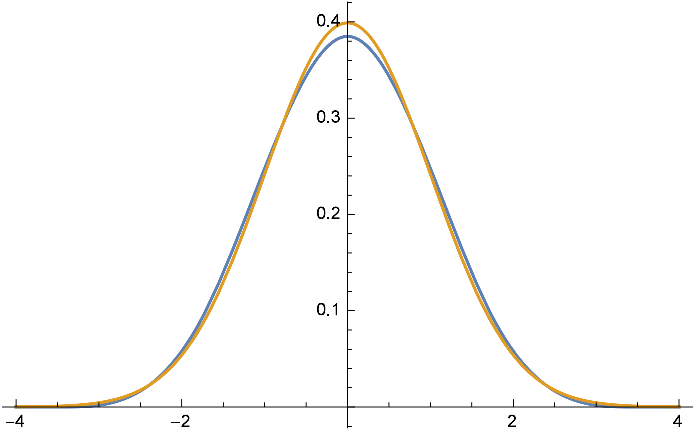
$$Y_2 = (X_1 + X_2)/\sigma_{X_1+X_2} \text{ vs } N(0,1).$$



Central Limit Theorem: Sums of Uniform Random Variables

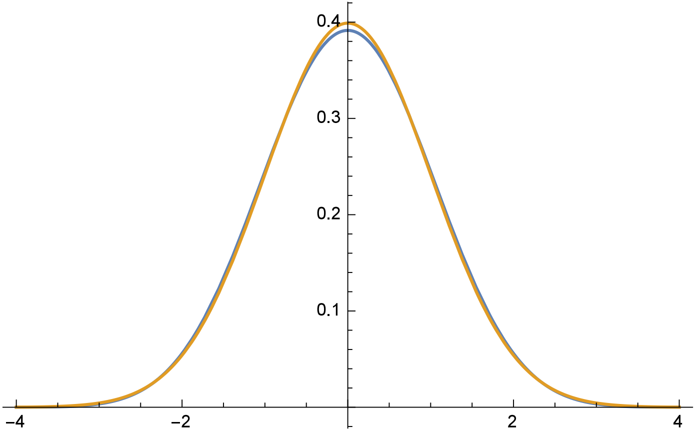
$X_i \sim \mathrm{Unif}(-1/2, 1/2)$ **(adjusted to mean 0, variance 1)**

$$Y_4 = (X_1 + X_2 + X_3 + X_4)/\sigma_{X_1+X_2+X_3+X_4} \text{ vs } N(0,1).$$



Central Limit Theorem: Sums of Uniform Random Variables $X_i \sim \mathrm{Unif}(-1/2, 1/2)$ (adjusted to mean 0, variance 1)

$$Y_8 = (X_1 + \cdots + X_8)/\sigma_{X_1+\cdots+X_8} \text{ vs } N(0,1).$$



Central Limit Theorem: Sums of Uniform Random Variables
$X_i \sim \mathrm{Unif}(-1/2, 1/2)$ (adjusted to mean 0, variance 1)

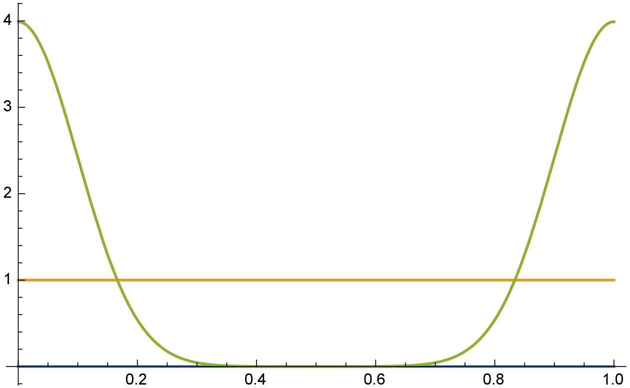
Density of $Y_4 = (X_1 + \cdots + X_4)/\sigma_{X_1+\cdots+X_4}$.

$$\frac{\begin{cases} \frac{1}{27}\left(18 + 9\sqrt{3}\, y - \sqrt{3}\, y^3\right) & y == 0 \\ \frac{1}{18}\left(12 - 6 y^2 - \sqrt{3}\, y^3\right) & -\sqrt{3} < y < 0 \\ \frac{1}{54}\left(72 - 36\sqrt{3}\, y + 18 y^2 - \sqrt{3}\, y^3\right) & \sqrt{3} < y < 2\sqrt{3} \\ \frac{1}{54}\left(18\sqrt{3}\, y - 18 y^2 + \sqrt{3}\, y^3\right) & y == \sqrt{3} \\ \frac{1}{18}\left(12 - 6 y^2 + \sqrt{3}\, y^3\right) & 0 < y < \sqrt{3} \\ \frac{1}{54}\left(72 + 36\sqrt{3}\, y + 18 y^2 + \sqrt{3}\, y^3\right) & -2\sqrt{3} < y \le -\sqrt{3} \\ 0 & \text{True} \end{cases}}{\sqrt{3}}$$

(Don't even think of asking to see Y_8's!)

Normal Distributions Mod 1

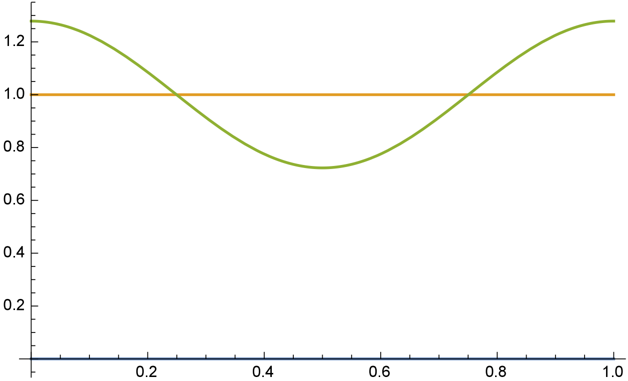
As $\sigma \to \infty$, $N(0, \sigma^2)$ mod 1 $\to$ Unif$(0, 1)$.



Variance is .01.

Normal Distributions Mod 1

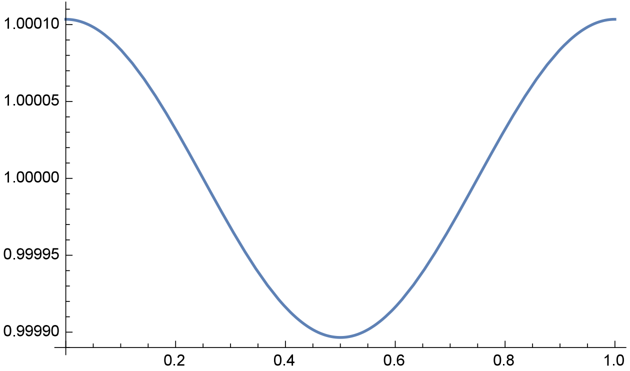
As $\sigma \to \infty$, $N(0, \sigma^2)$ mod 1 $\to$ Unif$(0, 1)$.



Variance is .1.

Normal Distributions Mod 1

As $\sigma \to \infty$, $N(0, \sigma^2)$ mod 1 $\to$ Unif$(0, 1)$.



Variance is .5.

Products and Benford's Law

Pavlovian Response: See a product, take a logarithm.

Products and Benford's Law

Pavlovian Response: See a product, take a logarithm.

$X_1, X_2, \dots$ nice, $W_N = X_1 \cdot X_2 \cdots X_N$.

Products and Benford's Law

Pavlovian Response: See a product, take a logarithm.

$X_1, X_2, \dots$ nice, $W_N = X_1 \cdot X_2 \cdots X_N$.

$Y_i = \log_{10} X_i$, $V_N := \log_{10} W_N$.

Products and Benford's Law

Pavlovian Response: See a product, take a logarithm.

$X_1, X_2, \dots$ nice, $W_N = X_1 \cdot X_2 \cdots X_N$.

$Y_i = \log_{10} X_i$, $V_N := \log_{10} W_N$.

$$V_N = \log_{10}(X_1 \cdot X_2 \cdots X_N)$$

Products and Benford's Law

Pavlovian Response: See a product, take a logarithm.

$X_1, X_2, \dots$ nice, $W_N = X_1 \cdot X_2 \cdots X_N$.

$Y_i = \log_{10} X_i$, $V_N := \log_{10} W_N$.

$$\begin{aligned} V_N &= \log_{10}(X_1 \cdot X_2 \cdots X_N) \\ &= \log_{10} X_1 + \log_{10} X_2 + \cdots + \log_{10} X_N \end{aligned}$$

Products and Benford's Law

Pavlovian Response: See a product, take a logarithm.

$X_1, X_2, \dots$ nice, $W_N = X_1 \cdot X_2 \cdots X_N$.

$Y_i = \log_{10} X_i$, $V_N := \log_{10} W_N$.

$$\begin{aligned} V_N &= \log_{10}(X_1 \cdot X_2 \cdots X_N) \\ &= \log_{10} X_1 + \log_{10} X_2 + \cdots + \log_{10} X_N \\ &= Y_1 + Y_2 + \cdots + Y_N. \end{aligned}$$

Need distribution of $V_N \bmod 1$, which by CLT becomes uniform, implying Benfordness!

Applications

Applications for the IRS: Detecting Fraud

A Tale of Two Steve Millers....

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.

Can you see the cat in the tree?

Transmitting Images

How to transmit an image?

- Have an $L \times W$ grid with LW pixels.
- Each pixel a triple: (Red, Green, Blue).
- Often each value in $\{0, 1, 2, 3, \dots, 2^n - 1\}$.
- $n = 8$ gives 256 choices for each, or 16,777,216 possibilities.

Steganography

Steganography: Concealing a message in another message: https://en.wikipedia.org/wiki/Steganography.

Steganography

Steganography: Concealing a message in another message: https://en.wikipedia.org/wiki/Steganography.

Take one of the colors, say red, a number from 0 to 255.

Write in binary: $r_7 2^7 + r_6 2^6 + \cdots + r_1 2 + r_0$.

If change just the last or last two digits, very minor change to image.

Can you see the cat in the tree?

Can you see the cat in the tree?

Benford Good Processes

- A. Kontorovich and S. J. Miller, *Benford's Law, values of L-functions and the* $3x+1$ *problem*, Acta Arithmetica **120** (2005), no. 3, 269–297.

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

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

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

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

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

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

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

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

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

Sketch of the proof (continued)

$$\begin{aligned}
&\sum_{\ell=-\infty}^{\infty} \mathbb{P}\left(a+\ell \leq \overrightarrow{Y}_{T,B} \leq b+\ell\right) \\
= &\sum_{|\ell| \leq Th(T)} \left[G_T(b+\ell) - G_T(a+\ell)\right] + o(1) \\
= &\int_a^b \sum_{|\ell| \leq Th(T)} \frac{1}{T} f\left(\frac{t+\ell}{T}\right) dt + \mathcal{E}(a,b,T) + o(1)
\end{aligned}$$

Sketch of the proof (continued)

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

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

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

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

Riemann Zeta Function (for real part of s greater than 1)

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{p \text{ prime}} \left(1 - \frac{1}{p^s}\right)^{-1}, \quad \mathrm{Re}(s) > 1.$$

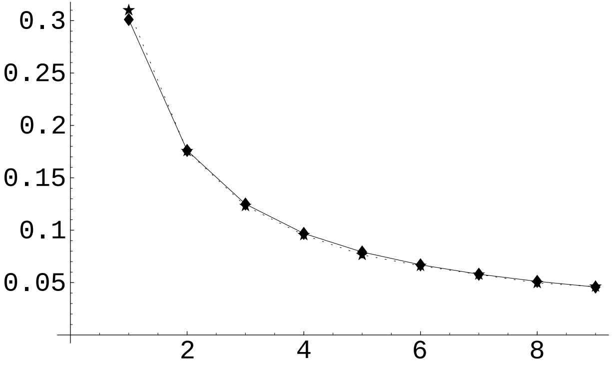
Geometric Series Formula: $(1-x)^{-1} = 1 + x + x^2 + \cdots$.
Unique Factorization: $n = p_1^{r_1} \cdots p_m^{r_m}$.

$$\prod_p \left(1 - \frac{1}{p^s}\right)^{-1} = \left[1 + \frac{1}{2^s} + \left(\frac{1}{2^s}\right)^2 + \cdots\right]\left[1 + \frac{1}{3^s} + \left(\frac{1}{3^s}\right)^2 + \cdots\right]\cdots$$

$$= \sum_n \frac{1}{n^s}.$$

Riemann Zeta Function

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



$3x + 1$ 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 $n = n(x)$, $T^n(x) = 1$.
- $7 \rightarrow_1 11 \rightarrow_1 17 \rightarrow_2 13$

$3x+1$ 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: $(k_1, \dots, k_m)$.

$3x+1$ 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.

$3x+1$ 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

Stick Decomposition

- T. Becker, D. Burt, T. C. Corcoran, A. Greaves-Tunnell, J. R. Iafrate, J. Jing, S. J. Miller, J. D. Porfilio, R. Ronan, J. Samranvedhya, F. W. Strauch and B. Talbut, *Benford's Law and Continuous Dependent Random Variables*, Annals of Physics **388** (2018), 350–381.
- J. Iafrate, S. J. Miller and F. W. Strauch, *Equipartitions and a distribution for numbers: A statistical model for Benford's law*, Physical Review E **91** (2015), no. 6, 062138 (6 pages).

Fixed Proportion Decomposition Process

Decomposition Process

1. Consider a stick of length $\mathcal{L}$.

Fixed Proportion Decomposition Process

Decomposition Process

1. Consider a stick of length $\mathcal{L}$.
2. Uniformly choose a proportion $p \in (0, 1)$.

Fixed Proportion Decomposition Process

Decomposition Process

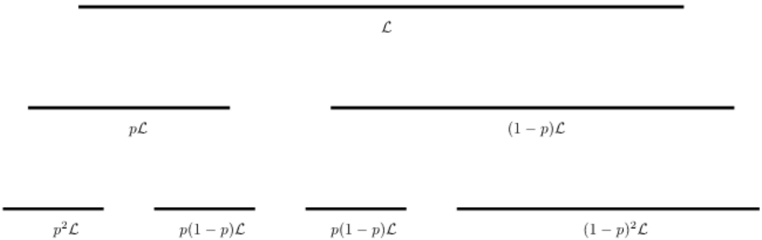
1. Consider a stick of length $\mathcal{L}$.
2. Uniformly choose a proportion $p \in (0, 1)$.
3. Break the stick into two pieces—lengths $p\mathcal{L}$ and $(1 - p)\mathcal{L}$.

Fixed Proportion Decomposition Process

Decomposition Process

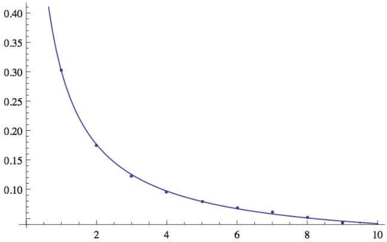
1. Consider a stick of length $\mathcal{L}$.
2. Uniformly choose a proportion $p \in (0, 1)$.
3. Break the stick into two pieces—lengths $p\mathcal{L}$ and $(1 - p)\mathcal{L}$.
4. Repeat N times (using the same proportion).

Fixed Proportion Decomposition Process

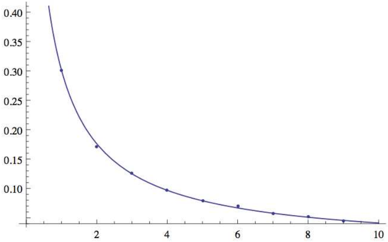


Fixed Proportion Conjecture (Joy Jing '13)

Conjecture: The above decomposition process is Benford as $N \to \infty$ for any $p \in (0,1)$, $p \neq \frac{1}{2}$.



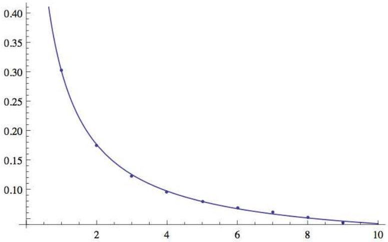
(B) $p = 0.51$ and $N = 10000$.



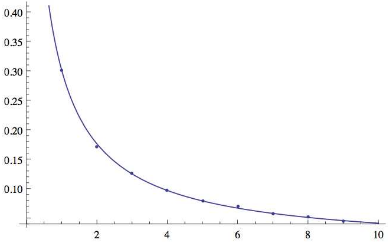
(B) $p = 0.99$ and $N = 50000$. Benford distribution overlaid.

Fixed Proportion Conjecture (Joy Jing '13)

Conjecture: The above decomposition process is Benford as $N \to \infty$ for any $p \in (0, 1)$, $p \neq \frac{1}{2}$.



(B) $p = 0.51$ and $N = 10000$.



(B) $p = 0.99$ and $N = 50000$. Benford distribution overlaid.

Counterexample (SMALL REU '13): $p = \frac{1}{11}$, $1 - p = \frac{10}{11}$.

Benford Analysis

At N^{th} level,

- 2^N sticks
- $N+1$ distinct lengths: write $p^{N-j}(1-p)^j$ as

$$p^N\left(\frac{1-p}{p}\right)^j,\ \ j \in \{0,\dots,N\},\ \text{have}\ \binom{N}{j}\ \text{times.}$$

Benford Analysis

At N^{th} level,

- 2^N sticks
- $N+1$ distinct lengths: write $p^{N-j}(1-p)^j$ as

$$p^N\left(\frac{1-p}{p}\right)^j, \ j \in \{0,\dots,N\}, \text{ have } \binom{N}{j} \text{ times.}$$

(Weighted) Geometric with ratio $\frac{1-p}{p} = 10^y$; behavior depends on irrationality of y!

Benford Analysis

At N^{th} level,

- 2^N sticks
- $N+1$ distinct lengths: write $p^{N-j}(1-p)^j$ as

$$p^N\left(\frac{1-p}{p}\right)^j,\ \ j \in \{0,\dots,N\},\ \text{have}\ \binom{N}{j}\ \text{times.}$$

(Weighted) Geometric with ratio $\frac{1-p}{p} = 10^y$;
behavior depends on irrationality of y!
Theorem: Benford if and only if y irrational.

Benford Analysis (cont)

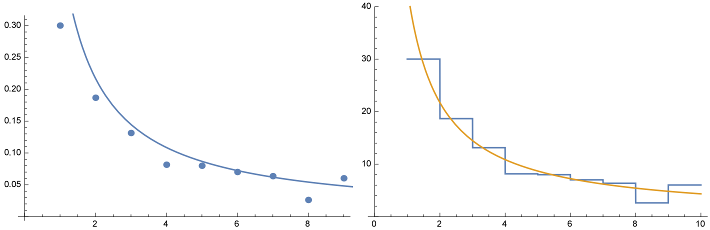
Say $\frac{1-p}{p} = 10^{r/q}$ for r, q integers.

All terms with index j mod q have same leading digit; probability index $j \mod q$ is

$$\begin{aligned}
\frac{1}{2^N}\left[\binom{N}{j}+\binom{N}{j+q}+\binom{N}{j+2q}+\cdots\right] &= \frac{1}{q}\sum_{s=0}^{q-1}\left(\cos\frac{\pi s}{q}\right)^N \cos\frac{\pi(N-2j)s}{q} \\
&= \frac{1}{q}\left(1+\sum_{s=1}^{q-1}\left(\cos\frac{\pi s}{q}\right)^N \cos\frac{\pi(N-2j)s}{q}\right) \\
&= \frac{1}{q}\left(1+\mathrm{Err}\left[(q-1)\left(\cos\frac{\pi}{q}\right)^N\right]\right),
\end{aligned}$$

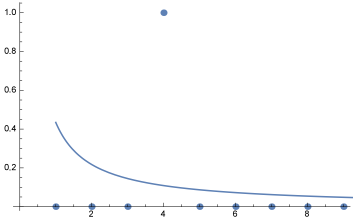
where $\mathrm{Err}[X]$ indicates an absolute error of size at most X

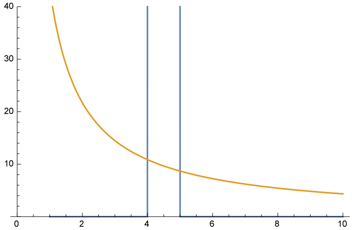
Examples



$p = 3/11$, 1000 levels; $y = \log_{10}(8/3) \notin \mathbb{Q}$ (irrational)

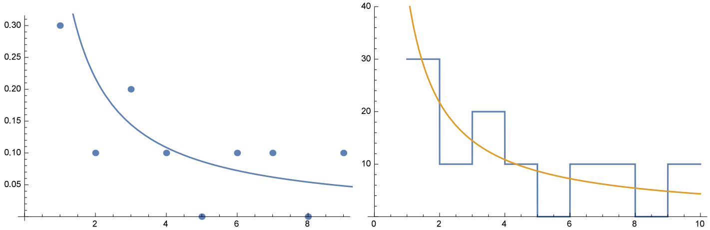
Examples





$p = 1/11$, 1000 levels; $y = 1 \in \mathbb{Q}$ (rational)

Examples



$p = 1/(1 + 10^{33/10})$, 1000 levels; $y = 33/10 \in \mathbb{Q}$ (rational)

Random Cuts

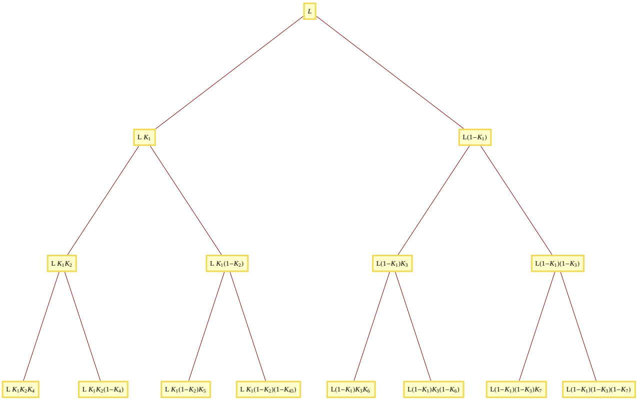


Figure: Unrestricted Decomposition: Breaking L into pieces, $N = 3$.

Conclusions and References

Conclusions and Future Investigations

- See many different systems exhibit Benford behavior.
- Ingredients of proofs (logarithms, equidistribution).
- Applications to fraud detection / data integrity.

- 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 J. 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. `http://www.jstor.org/stable/984802?seq=1#page_scan_tab_contents`.

- A. Berger, L. A. Bunimovich and T. Hill, *One-dimensional dynamical systems and Benford's Law*, Trans. AMS **357** (2005), no. 1, 197–219. http://www.ams.org/journals/tran/2005-357-01/S0002-9947-04-03455-5/.

- A. Berger and T. Hill, *Newton's method obeys Benford's law*, The Amer. Math. Monthly **114** (2007), no. 7, 588-601. http://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1058&context=rgp_rsr.

- A. Berger and T. Hill, *Benford on-line bibliography*, http://www.benfordonline.net/.

- J. Boyle, *An application of Fourier series to the most significant digit problem* Amer. Math. Monthly **101** (1994), 879–886. http://www.jstor.org/stable/2975136?seq=1#page_scan_tab_contents.

- 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. http://statweb.stanford.edu/~cgates/PERSI/papers/digits.pdf.

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. https://archive.org/details/bstj49-8-1609.

T. Hill, *The first-digit phenomenon*, American Scientist **86** (1996), 358–363. http://www.americanscientist.org/issues/feature/1998/4/the-first-digit-phenomenon/99999.

T. Hill, *A statistical derivation of the significant-digit law*, Statistical Science **10** (1996), 354–363. https://projecteuclid.org/euclid.ss/1177009869.

- 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. Kang, A. Kruckman, J. Kudo & 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.

- E. Janvresse and T. de la Rue, *From uniform distribution to Benford's law*, Journal of Applied Probability **41** (2004) no. 4, 1203–1210. http://www.jstor.org/stable/4141393?seq=1#page_scan_tab_contents.

- 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/pdf/math/0412003.pdf.

- 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/0509175.pdf.
- 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. http://www.jstor.org/stable/2684926?seq=1#page_scan_tab_contents.

- 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, *Benford's Law: Theory and Applications*, Princeton University Press, 2015. `http://web.williams.edu/Mathematics/sjmiller/public_html/benford/`.

- 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/0607686v2`.

- 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. `http://arxiv.org/pdf/math/0601344v5.pdf`.

- S. J. Miller and R. Takloo-Bighash, *An Invitation to Modern Number Theory*, Princeton University Press, Princeton, NJ, 2006. http://web.williams.edu/Mathematics/sjmiller/public_html/book/index.html.

- S. Newcomb, *Note on the frequency of use of the different digits in natural numbers*, Amer. J. Math. **4** (1881), 39-40. http://www.jstor.org/stable/2369148?seq=1#page_scan_tab_contents.

- 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. http://link.springer.com/article/10.1007%2Fs11004-007-9109-5?LI=true.

- 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. http://accounting.uwaterloo.ca/uwcisa/symposiums/symposium_2007/AdvancedBenfordsLaw7.pdf.

- 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. http://projecteuclid.org/download/pdf_1/euclid.aoms/1177704862.

- 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. http://www.jstor.org/stable/2946226?seq=1#page_scan_tab_contents.
- K. Stromberg, *Probabilities on a compact group*, Trans. Amer. Math. Soc. **94** (1960), 295–309. http://www.jstor.org/stable/1993313?seq=1#page_scan_tab_contents.
- P. R. Turner, *The distribution of leading significant digits*, IMA J. Numer. Anal. **2** (1982), no. 4, 407–412.

Products of Random Variables

- 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.

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

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

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

Modulo 1 Central Limit Theorem

Theorem (M– and Nigrini 2007)

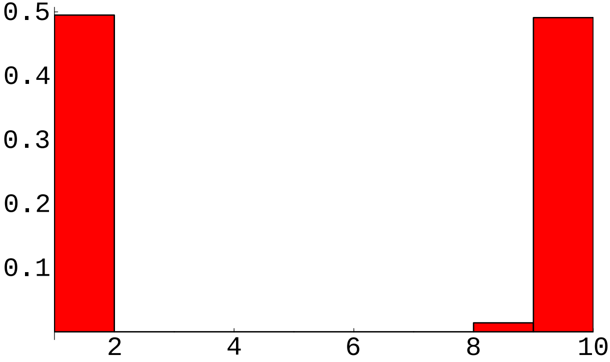
$\{Y_m\}$ *independent continuous random variables on* $[0,1)$ *(not necc. i.i.d.), densities* $\{g_m\}$*.* $Y_1 + \cdots + Y_M \bmod 1$ *converges to the uniform distribution as* $M \to \infty$ *in* $L^1([0,1])$ *if and only if for all* $n \neq 0$, $\lim_{M\to\infty} \widehat{g}_1(n) \cdots \widehat{g_M}(n) = 0$*.*

◇ 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 $|x - 1/8| \le 1/2m$ (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

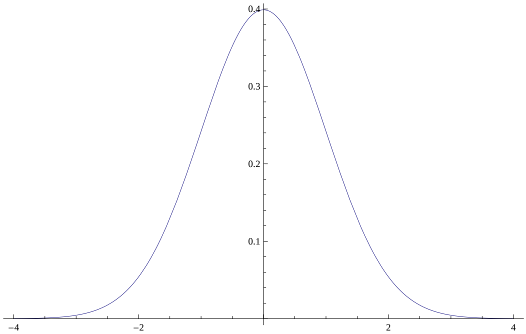


Figure: Plot of normal (mean 0, stdev 1).

Proof under stronger conditions

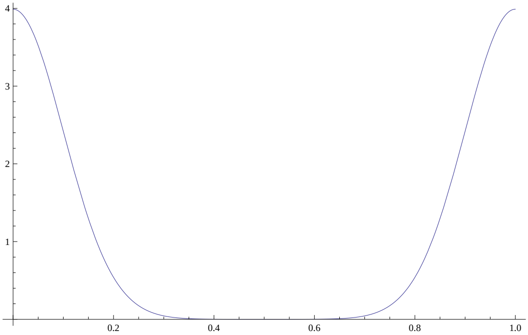


Figure: Plot of normal (mean 0, stdev .1) modulo 1.

Proof under stronger conditions

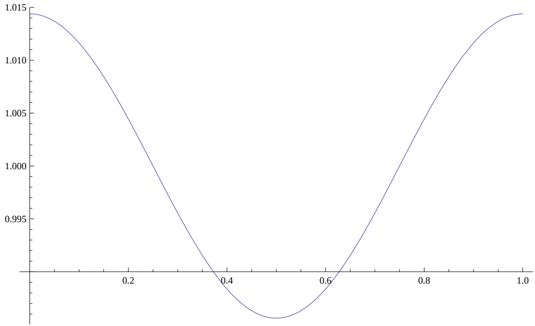


Figure: Plot of normal (mean 0, stdev .5) modulo 1.

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

$$\begin{aligned}
&\frac{1}{\sqrt{N}} \sum_{|n| \le N^{5/4}} \int_{x=a}^{b} e^{-\pi (x+n)^2/N} dx \\
&= \frac{1}{\sqrt{N}} \sum_{|n| \le N^{5/4}} \int_{x=a}^{b} \left[e^{-\pi n^2/N} + O\left(\frac{\max(1,|n|)}{N} e^{-n^2/N} \right) \right] dx \\
&= \frac{b-a}{\sqrt{N}} \sum_{|n| \le N^{5/4}} e^{-\pi n^2/N} + O\left(\frac{1}{N} \sum_{n=0}^{N^{5/4}} \frac{n+1}{\sqrt{N}} e^{-\pi (n/\sqrt{N})^2} \right) \\
&= \frac{b-a}{\sqrt{N}} \sum_{|n| \le N^{5/4}} e^{-\pi n^2/N} + O\left(\frac{1}{N} \int_{w=0}^{N^{3/4}} (w+1) e^{-\pi w^2} \sqrt{N} dw \right) \\
&= \frac{b-a}{\sqrt{N}} \sum_{|n| \le N^{5/4}} e^{-\pi n^2/N} + O\left(N^{-1/2} \right).
\end{aligned}$$

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:

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

- Fejér series $T_N f(x)$ equals

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

- 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

- 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.

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 $\Xi_1 \cdot \Xi_2$:

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

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

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

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

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: $Y_n = \log_B \Xi_n$,
$|\text{Prob}(Y_n \bmod 1 \in [a,b]) - (b-a)| \le$

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

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*

$$\left|\mathrm{Prob}(Y_n \bmod 1 \in [a,b]) - (b+a)\right| \leq$$

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

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

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

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

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

Bounds on the error

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

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