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/

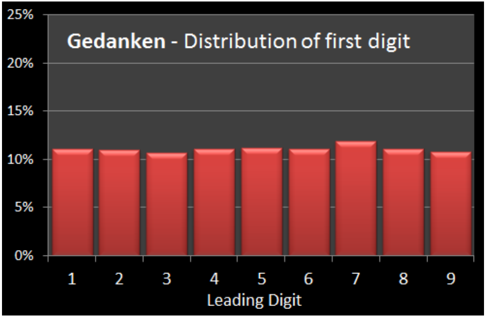
Carnegie Mellon University, March 25, 2/12/15
Winona State University, March 30, 2/12/15

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?

Interesting Question

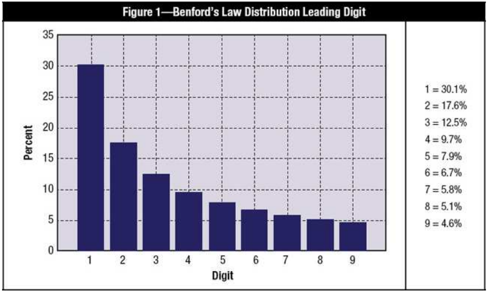
Motivating Question: For a nice data set, such as the Fibonacci numbers, stock prices, street addresses of college employees and students, ..., what percent of the leading digits are 1?



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

Interesting Question

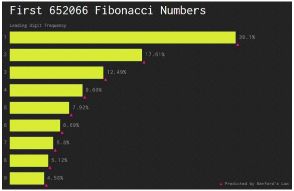
Motivating Question: For a nice data set, such as the Fibonacci numbers, stock prices, street addresses of college employees and students, ..., what percent of the leading digits are 1?



Answer: Benford's law!

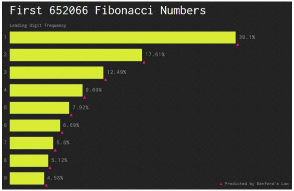
Examples with First Digit Bias

Fibonacci numbers

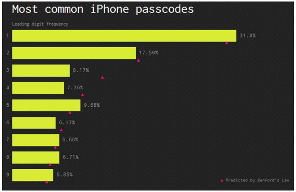


Examples with First Digit Bias

Fibonacci numbers

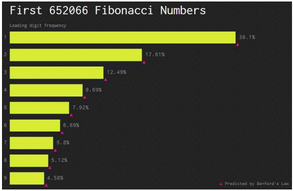


Most common iPhone passcodes

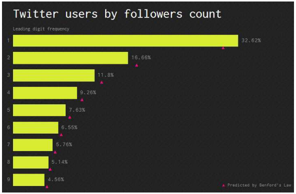


Examples with First Digit Bias

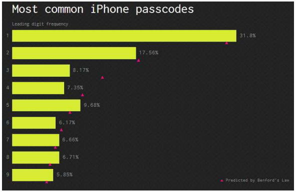
Fibonacci numbers



Twitter users by # followers

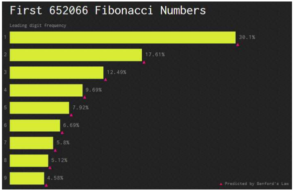


Most common iPhone passcodes

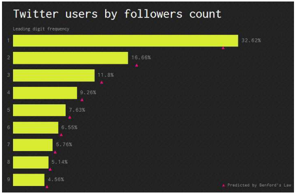


Examples with First Digit Bias

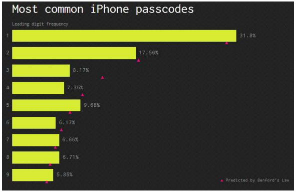
Fibonacci numbers



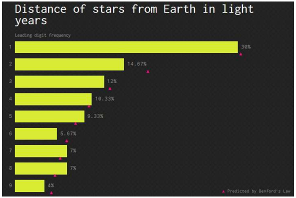
Twitter users by # followers



Most common iPhone passcodes



Distance of stars from Earth



Summary

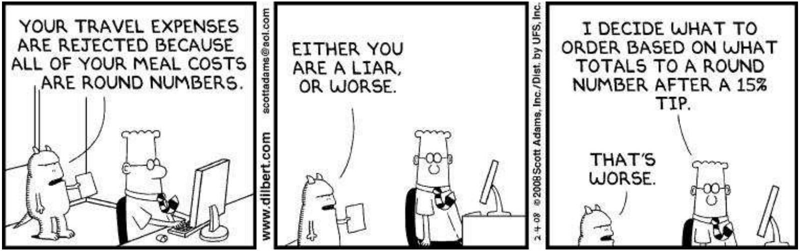
- Explain Benford's Law.
- Discuss examples and applications.
- Sketch proofs.
- Describe open problems.

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

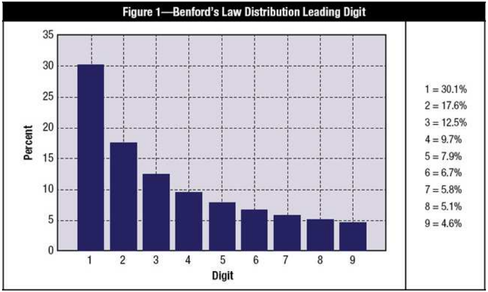
- Analyzing round-off errors.
- Determining the optimal way to store numbers.
- Detecting tax and image fraud, and data integrity.

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, $1 \le S_{10}(x) < 10$.

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

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

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 $b-a$:

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

Equidistribution and Benford's Law

Equidistribution

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

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

- 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

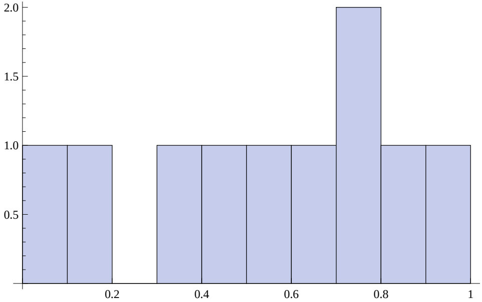
Equidistribution

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

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

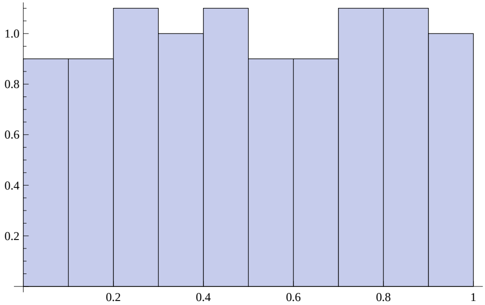
- Thm: $\beta \notin \mathbb{Q}$, $n\beta$ is equidistributed mod 1.
- Examples: $\log_{10} 2, \log_{10}\left(\frac{1+\sqrt{5}}{2}\right) \notin \mathbb{Q}$.
  Proof: if rational: $2 = 10^{p/q}$.
  Thus $2^q = 10^p$ or $2^{q-p} = 5^p$, impossible.

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



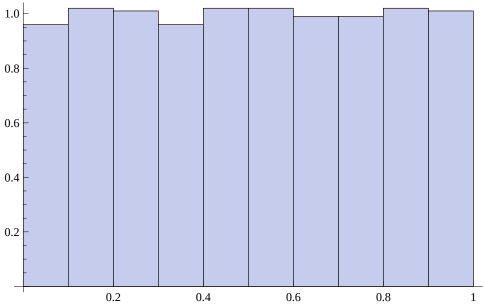
$n\sqrt{\pi} \bmod 1$ for $n \leq 10$

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



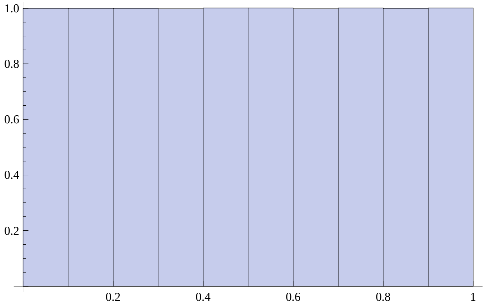
$n\sqrt{\pi}$ mod 1 for $n \le 100$

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



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

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



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

Logarithms and Benford's Law

Fundamental Equivalence

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

Logarithms and Benford's Law

Fundamental Equivalence

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

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

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

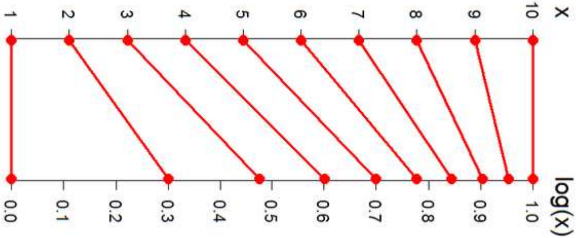
Logarithms and Benford's Law

Fundamental Equivalence

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

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

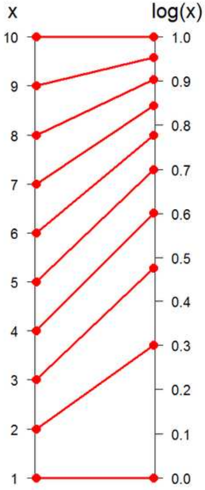
$$\log_{10} x = \log_{10} S_{10}(x) + k = \log_{10} S_{10} x \bmod 1.$$



Logarithms and Benford's Law

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

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



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

Examples

- Fibonacci numbers are Benford base 10.

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

- Most linear recurrence relations Benford:
  - ◇ $a_{n+1} = 2a_n - a_{n-1}$
  - ◇ take $a_0 = a_1 = 1$ or $a_0 = 0$, $a_1 = 1$.

Digits of 2^n

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

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

Digits of 2^n

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

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

Digits of 2^n

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

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

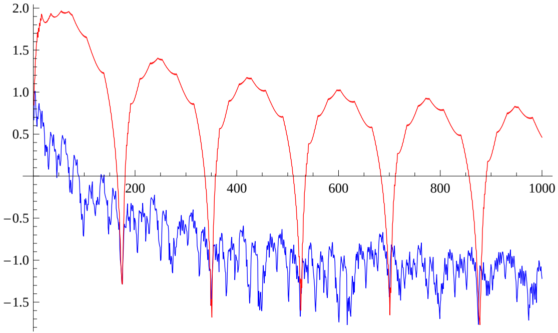
Logarithms and Benford's Law

χ^2 values for α^n, $1 \le n \le N$ (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

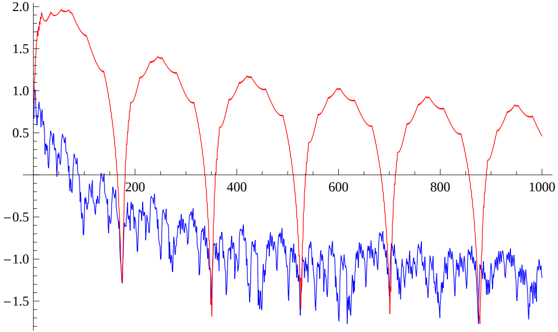
Logarithms and Benford's Law: Base 10 (5%: $\log(\chi^2) \approx 2.74$)

$\log(\chi^2)$ 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 $\pi^{175} \approx 1.0028 \cdot 10^{87}$.



Why Benford's Law?

Streets

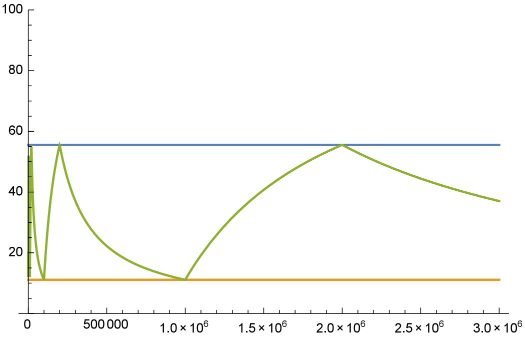
Not all data sets satisfy Benford's Law.

- 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.

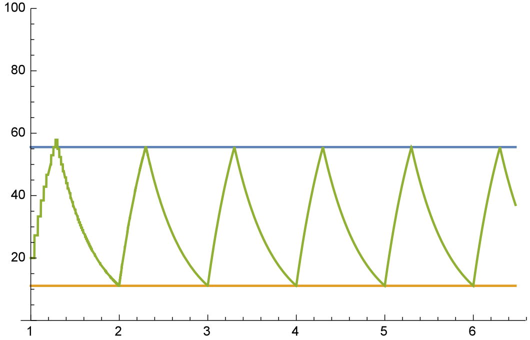


Probability first digit 1 versus street length L.

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.

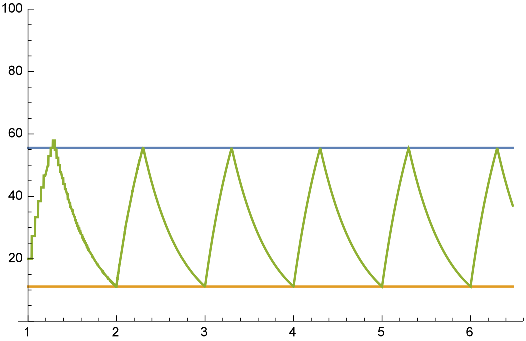


Probability first digit 1 versus log(street length L).

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.



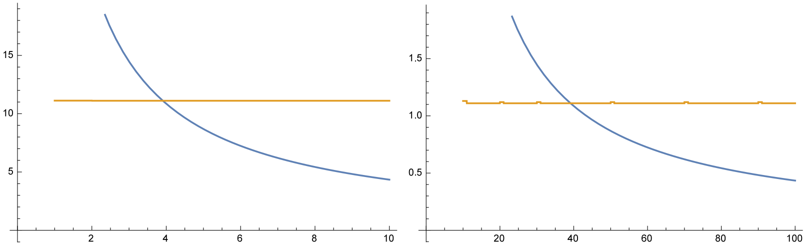
Probability first digit 1 versus log(street length L).

What if we have many streets of different lengths?

Amalgamating Streets

All houses: 1000 Streets,

each from 1 to 10000.

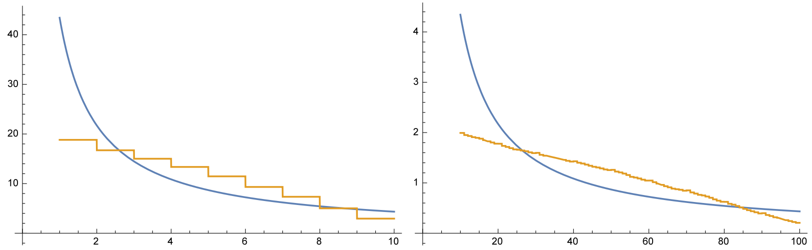


First digit and first two digits vs Benford.

Amalgamating Streets

All houses: 1000 Streets,

each from 1 to rand(10000).

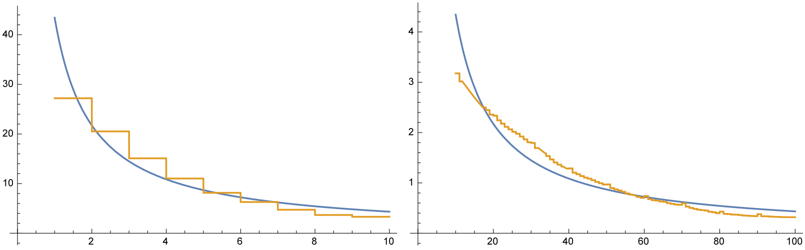


First digit and first two digits vs Benford.

Amalgamating Streets

All houses: 1000 Streets,

each 1 to rand(rand(10000)).



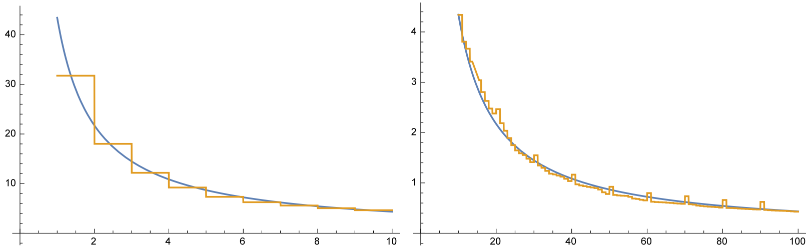
First digit and first two digits vs Benford.

Conclusion: More processes, closer to Benford.

Amalgamating Streets

All houses: 1000 Streets,

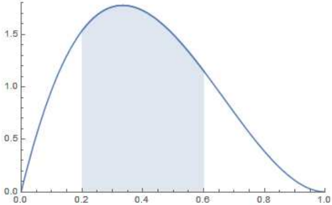
each 1 to rand(rand(rand(10000))).



First digit and first two digits vs Benford.

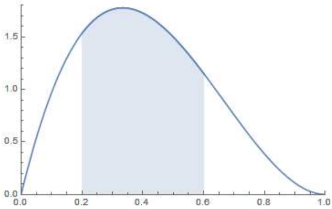
Conclusion: More processes, closer to Benford.

Probability Review



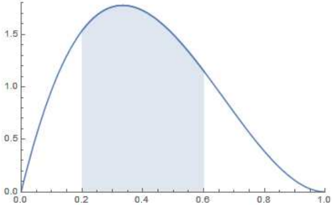
- Let X be random variable with density $p(x)$:
  - ◇ $p(x) \geq 0$; $\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$;
  - ◇ $\mathrm{Prob}\,(a \leq X \leq b) = \int_a^b p(x)\mathrm{d}x$.

Probability Review



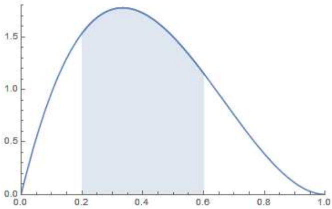
- Let X be random variable with density $p(x)$:
  - ⋄ $p(x) \geq 0$; $\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$;
  - ⋄ $\mathrm{Prob}\,(a \leq X \leq b) = \int_a^b p(x)\mathrm{d}x$.
- Mean $\mu = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x$.

Probability Review



- Let X be random variable with density $p(x)$:
  - $\diamond$ $p(x) \geq 0$; $\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$;
  - $\diamond$ $\mathrm{Prob}\,(a \leq X \leq b) = \int_a^b p(x)\mathrm{d}x$.
- Mean $\mu = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x$.
- Variance $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\mathrm{d}x$.

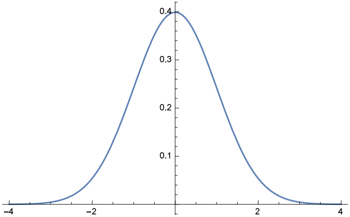
Probability Review



- Let X be random variable with density $p(x)$:
  $\diamond$ $p(x) \geq 0$; $\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$;
  $\diamond$ $\mathrm{Prob}\,(a \leq X \leq b) = \int_a^b p(x)\mathrm{d}x$.
- Mean $\mu = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x$.
- 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

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



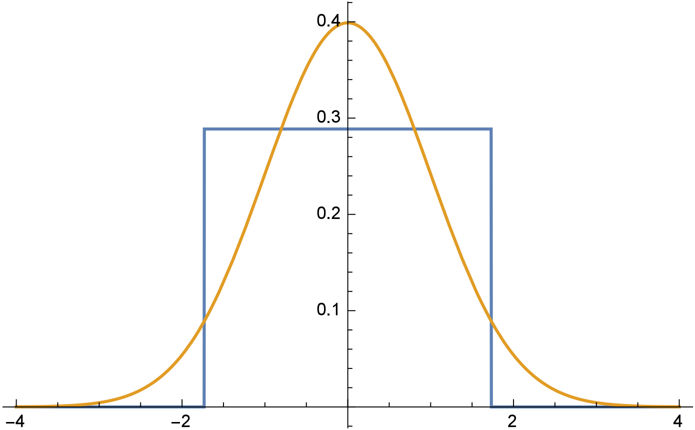
Theorem

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

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

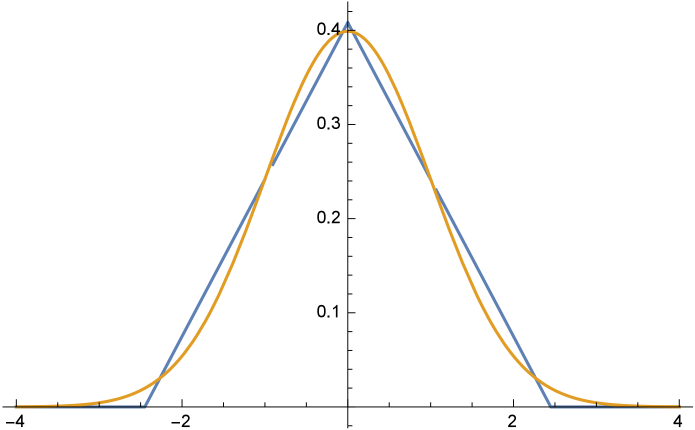
Central Limit Theorem: Sums of Uniform Random Variables $X_i \sim \mathrm{Unif}(-1/2, 1/2)$ (adjusted to mean 0, variance 1)

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



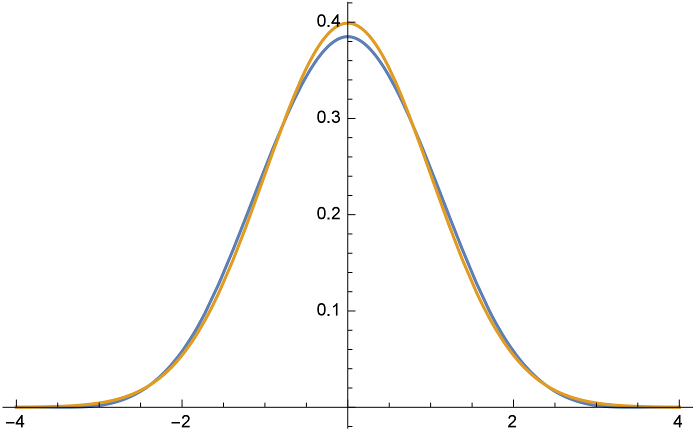
Central Limit Theorem: Sums of Uniform Random Variables
$X_i \sim \mathrm{Unif}(-1/2, 1/2)$ (adjusted to mean 0, variance 1)

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



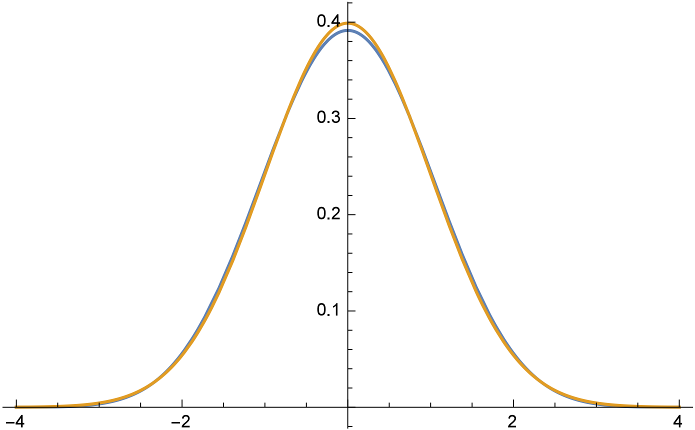
Central Limit Theorem: Sums of Uniform Random Variables
$X_i \sim \mathrm{Unif}(-1/2, 1/2)$ (adjusted to mean 0, variance 1)

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



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)

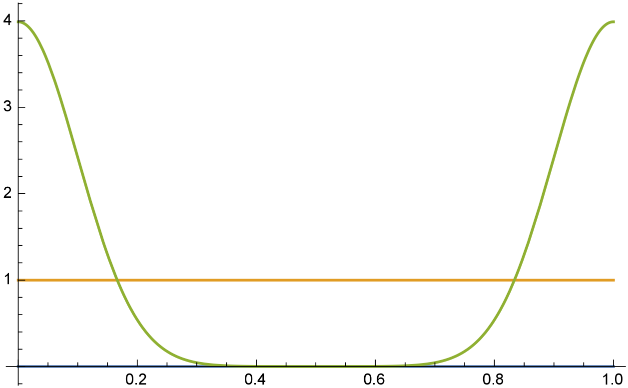
Density of $Y_4 = (X_1 + \cdots + X_4)/\sigma_{X_1+\cdots+X_4}$.

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

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

Normal Distributions Mod 1

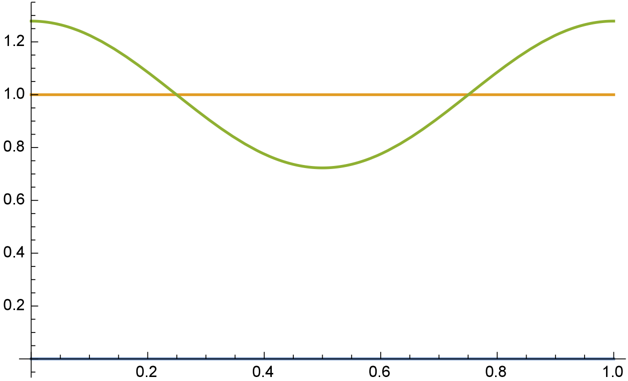
As $\sigma \to \infty$, $N(0, \sigma^2)$ mod 1 $\to \mathrm{Unif}(0, 1)$.



Variance is .01.

Normal Distributions Mod 1

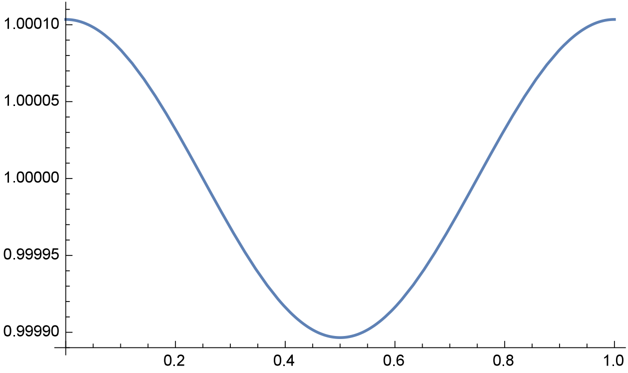
As $\sigma \to \infty$, $N(0, \sigma^2)$ mod 1 $\to \mathrm{Unif}(0, 1)$.



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.

$X_1, X_2, \ldots$ 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, $W_N = X_1 \cdot X_2 \cdots X_N$.

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

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

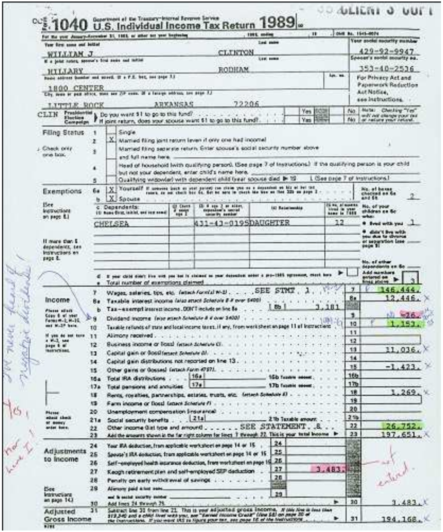
Need distribution of V_N mod 1, which by CLT becomes uniform, implying Benfordness!

Applications

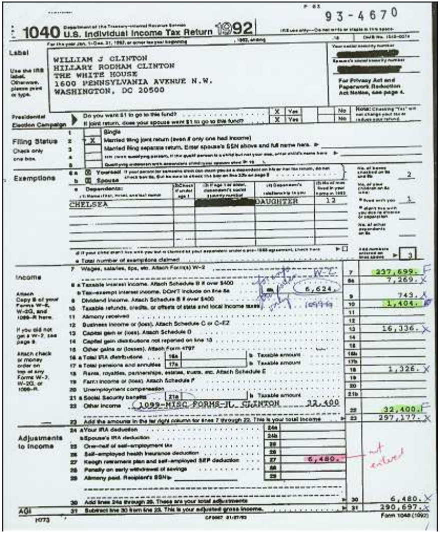
Applications for the IRS: Detecting Fraud

A Tale of Two Steve Millers....

Applications for the IRS: Detecting Fraud


1040 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 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.
23 197,651.
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
31 194,168.

Applications for the IRS: Detecting Fraud



93-4670

1040 U.S. Individual Income Tax Return 1992

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]

Presidential Election Campaign
Filing Status
Exemptions

CHELSEA [redacted] DAUGHTER 12

Income

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

Adjustments to Income

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

AGI

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

Form 1040 (1992)

Detecting Fraud

Bank Fraud

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

Detecting Fraud

Bank Fraud

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

Detecting Fraud

Bank Fraud

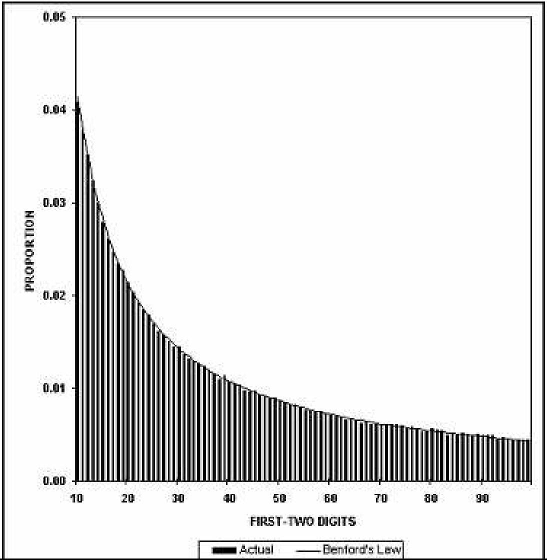
- Audit of a bank revealed huge spike of numbers starting with 48 and 49, most due to one person.

Detecting Fraud

Bank Fraud

- Audit of a bank revealed huge spike of numbers starting with 48 and 49, most due to one person.
- Write-off limit of $5,000. Officer had friends applying for credit cards, ran up balances just under $5,000 then he would write the debts off.

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



Election Fraud: Iran 2009

Numerous questions 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...).

The Riemann Zeta Function $\zeta(s)$ and Benford's Law

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

$$\zeta(s) \;=\; \sum_{n=1}^{\infty} \frac{1}{n^s}$$

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

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

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: $n = p_1^{r_1} \cdots p_m^{r_m}$.

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 (cont)

$$\zeta(s) = \sum_n \frac{1}{n^s} = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1}, \quad \mathrm{Re}(s) > 1$$

$$\pi(x) = \#\{p : p \text{ is prime}, p \leq x\}$$

Properties of $\zeta(s)$ and Primes:

Riemann Zeta Function (cont)

$$\zeta(s) = \sum_n \frac{1}{n^s} = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1}, \quad \mathrm{Re}(s) > 1$$

$$\pi(x) = \#\{p : p \text{ is prime}, p \le x\}$$

Properties of $\zeta(s)$ and Primes:

- $\lim_{s \to 1^+} \zeta(s) = \infty$, $\pi(x) \to \infty$.

Riemann Zeta Function (cont)

$$\zeta(s) = \sum_n \frac{1}{n^s} = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1}, \quad \mathrm{Re}(s) > 1$$

$$\pi(x) = \#\{p : p \text{ is prime}, p \leq x\}$$

Properties of $\zeta(s)$ and Primes:

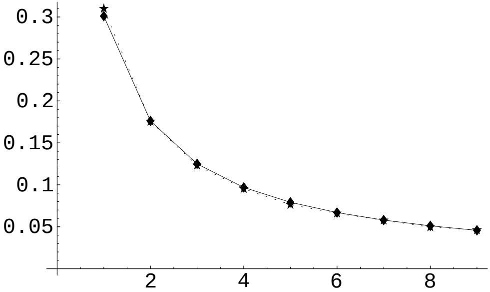
- $\lim_{s\to 1^+} \zeta(s) = \infty$, $\pi(x) \to \infty$.
- $\zeta(2) = \frac{\pi^2}{6}$, $\pi(x) \to \infty$.

The Riemann Zeta Function and Benford's Law

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

The Riemann Zeta Function and Benford's Law

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



First digits of $\left|\zeta\left(\frac{1}{2}+\mathrm{i}\frac{k}{4}\right)\right|$ versus Benford's law.

Proof Sketch: ‘Good’ L-Functions

We say an L-function is *good* if:

- Euler product:

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

- $L(s,f)$ has a meromorphic continuation to $\mathbb{C}$, is of finite order, and has at most finitely many poles (all on the line $\mathrm{Re}(s)=1$).
- Functional equation:

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

where $\omega \in \mathbb{R}$ and

$$G(s) \;=\; Q^s \prod_{i=1}^{h} \Gamma(\lambda_i s + \mu_i)$$

with $Q, \lambda_i > 0$ and $\mathrm{Re}(\mu_i) \geq 0$.

Proof Sketch: ‘Good’ L-Functions (cont)

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

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

- The $\alpha_{f,j}(p)$ are (Ramanujan-Petersson) tempered: $|\alpha_{f,j}(p)| \leq 1$.
- If $N(\sigma, T)$ is the number of zeros ρ of $L(s)$ with $\mathrm{Re}(\rho) \geq \sigma$ and $\mathrm{Im}(\rho) \in [0, T]$, then for some $\beta > 0$ we have

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

Known in some cases, such as $\zeta(s)$ and Hecke cuspidal forms of full level and even weight $k > 0$.

Log-Normal Law (Hejhal, Laurinčikas, Selberg)

Log-Normal Law

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

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

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

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

Result: Values of L-functions and Benford's Law

Theorem (Kontorovich and M–, 2005)

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

Ingredients

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

Results: Explicit L-Function Statement

Theorem (Kontorovich-Miller '05)

Let $L(s,f)$ be a good L-function. Fix a $\delta \in (0,1)$. For each T, let $\sigma_T = \frac{1}{2} + \frac{1}{\log^\delta T}$. Then as $T \to \infty$

$$\frac{\mu\left\{t \in [T, 2T] :\ M_B\left(|L(\sigma_T + \mathrm{i}t, f)|\right) \le \tau\right\}}{T} \to \log_B \tau$$

Thus the values of the L-function satisfy Benford's Law in the limit for any base B.

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 $n = n(x)$, $T^n(x) = 1$.

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

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

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 $\log(3/4) < 1$.

$3x + 1$ 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.

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, $1/2$):

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

Sketch of the proof of Benfordness

- 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 $\kappa < \infty$:

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

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.

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 $\kappa < 1.2 \cdot 10^{602}$:

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

Quantified Equidistribution

Theorem (Erdös-Turan)

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

There is a C such that for all m:

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

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

Stick Decomposition

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

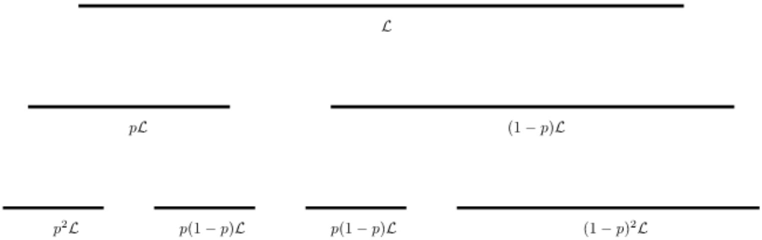
1. Consider a stick of length $\mathcal{L}$.
2. Uniformly choose a proportion $p \in (0, 1)$.
3. Break the stick into two pieces—lengths $p\mathcal{L}$ and $(1-p)\mathcal{L}$.

Fixed Proportion Decomposition Process

Decomposition Process

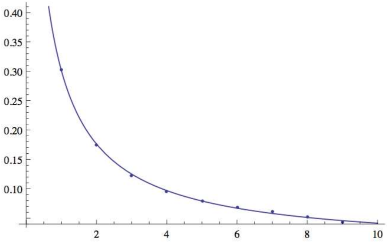
1. Consider a stick of length $\mathcal{L}$.
2. Uniformly choose a proportion $p \in (0, 1)$.
3. Break the stick into two pieces—lengths $p\mathcal{L}$ and $(1 - p)\mathcal{L}$.
4. Repeat N times (using the same proportion).

Fixed Proportion Decomposition Process

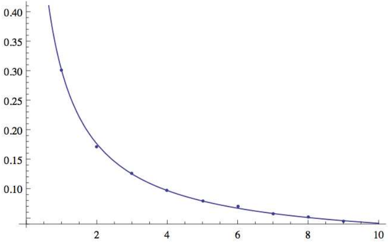


Fixed Proportion Conjecture (Joy Jing '13)

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



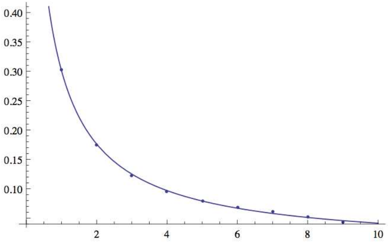
(B) $p = 0.51$ and $N = 10000$.



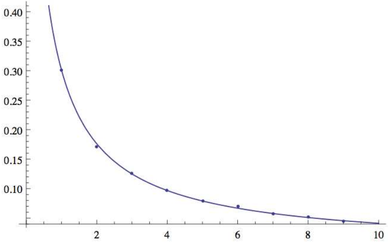
(B) $p = 0.99$ and $N = 50000$. Benford distribution overlaid.

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

Benford Analysis

At N^{th} level,

- 2^N sticks
- $N+1$ distinct lengths:

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

Benford Analysis

At N^{th} level,

- 2^N sticks
- $N+1$ distinct lengths:

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

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

Benford Analysis

At N^{th} level,

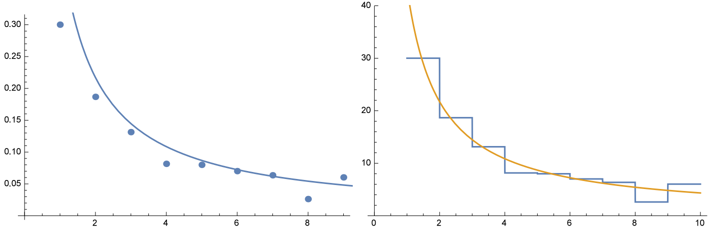
- 2^N sticks
- $N+1$ distinct lengths:

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

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

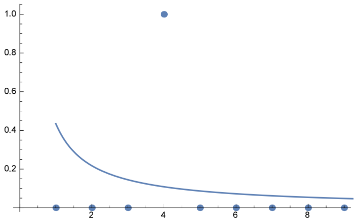
Theorem: Benford if and only if y irrational.

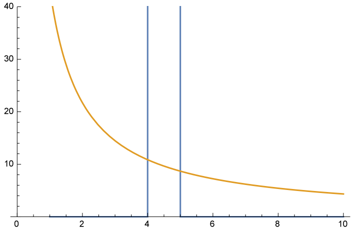
Examples



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

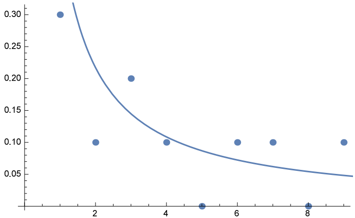
Examples

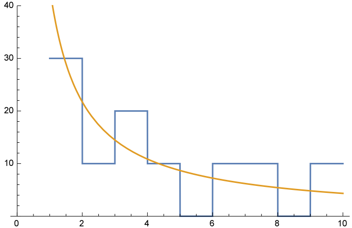




$p = 1/11$, 1000 levels; $y = 1 \in \mathbb{Q}$ (rational)

Examples





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

Conclusions

Current / Future Investigations

- Develop more sophisticated tests for fraud.

- Study digits of other systems.
  - ◇ Break rod of fixed length a variable number of times.
  - ◇ Break rods of variable length a variable number of times.
  - ◇ Break rods of variable length, each piece then breaks with given probability.
  - ◇ Break rods of variable integer length, each piece breaks until is a prime, or a square, ...

Conclusions and Future Investigations

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

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 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, in press, expected publication date 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.