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

Steven J. Miller, Williams College

sjm1@williams.edu
http://www.williams.edu/Mathematics/sjmiller/

Hampshire College, July 14, 2011

Interesting Question

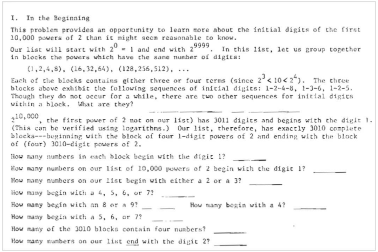
Interesting Question

For a nice data set, such as the Fibonacci numbers, what percent of the leading digits are 1?

Interesting Question

Interesting Question

For a nice data set, such as the Fibonacci numbers, what percent of the leading digits are 1?



I. In the Beginning

This problem provides an opportunity to learn more about the initial digits of the first 10,000 powers of 2 than it might seem reasonable to know.

Our list will start with $2^0 = 1$ and end with 2^{9999}. In this list, let us group together in blocks the powers which have the same number of digits:

(1,2,4,8), (16,32,64), (128,256,512), ...

Each of the blocks contains either three or four terms (since $2^3 < 10 < 2^4$). The three blocks above exhibit the following sequences of initial digits: 1-2-4-8, 1-3-6, 1-2-5. Though they do not occur for a while, there are two other sequences for initial digits within a block. What are they? ______ ______

$2^{10,000}$, the first power of 2 not on our list) has 3011 digits and begins with the digit 1. (This can be verified using logarithms.) Our list, therefore, has exactly 3010 complete blocks---beginning with the block of four 1-digit powers of 2 and ending with the block of (four) 3010-digit powers of 2.

How many numbers in each block begin with the digit 1? ______

How many numbers on our list of 10,000 powers of 2 begin with the digit 1? ______

How many numbers on our list begin with either a 2 or a 3? ______

How many begin with a 4, 5, 6, or 7? ______

How many begin with an 8 or a 9? ______ How many begin with a 4? ______

How many begin with a 5, 6, or 7? ______

How many of the 3010 blocks contain four numbers? ______

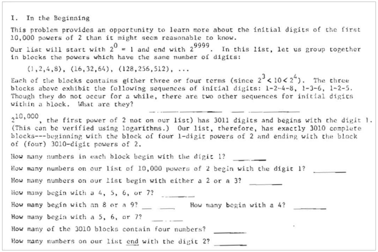
How many numbers on our list end with the digit 2? ______

Plausible answers:

Interesting Question

Interesting Question

For a nice data set, such as the Fibonacci numbers, what percent of the leading digits are 1?



I. In the Beginning

This problem provides an opportunity to learn more about the initial digits of the first 10,000 powers of 2 than it might seem reasonable to know.

Our list will start with $2^0 = 1$ and end with 2^{9999}. In this list, let us group together in blocks the powers which have the same number of digits:

(1,2,4,8), (16,32,64), (128,256,512), ...

Each of the blocks contains either three or four terms (since $2^3 < 10 < 2^4$). The three blocks above exhibit the following sequences of initial digits: 1-2-4-8, 1-3-6, 1-2-5. Though they do not occur for a while, there are two other sequences for initial digits within a block. What are they? ______ ______

$2^{10,000}$, the first power of 2 not on our list) has 3011 digits and begins with the digit 1. (This can be verified using logarithms.) Our list, therefore, has exactly 3010 complete blocks---beginning with the block of four 1-digit powers of 2 and ending with the block of (four) 3010-digit powers of 2.

How many numbers in each block begin with the digit 1? ______

How many numbers on our list of 10,000 powers of 2 begin with the digit 1? ______

How many numbers on our list begin with either a 2 or a 3? ______

How many begin with a 4, 5, 6, or 7? ______

How many begin with an 8 or a 9? ______ How many begin with a 4? ______

How many begin with a 5, 6, or 7? ______

How many of the 3010 blocks contain four numbers? ______

How many numbers on our list end with the digit 2? ______

Plausible answers: 10%

Interesting Question

Interesting Question

For a nice data set, such as the Fibonacci numbers, what percent of the leading digits are 1?

I. In the Beginning

This problem provides an opportunity to learn more about the initial digits of the first 10,000 powers of 2 than it might seem reasonable to know.

Our list will start with $2^0 = 1$ and end with 2^{9999}. In this list, let us group together in blocks the powers which have the same number of digits:

(1,2,4,8), (16,32,64), (128,256,512), ...

Each of the blocks contains either three or four terms (since $2^3 < 10 < 2^4$). The three blocks above exhibit the following sequences of initial digits: 1-2-4-8, 1-3-6, 1-2-5. Though they do not occur for a while, there are two other sequences for initial digits within a block. What are they? ______ ______

$2^{10,000}$, the first power of 2 not on our list) has 3011 digits and begins with the digit 1. (This can be verified using logarithms.) Our list, therefore, has exactly 3010 complete blocks---beginning with the block of four 1-digit powers of 2 and ending with the block of (four) 3010-digit powers of 2.

How many numbers in each block begin with the digit 1? ______

How many numbers on our list of 10,000 powers of 2 begin with the digit 1? ______

How many numbers on our list begin with either a 2 or a 3? ______

How many begin with a 4, 5, 6, or 7? ______

How many begin with an 8 or a 9? ______ How many begin with a 4? ______

How many begin with a 5, 6, or 7? ______

How many of the 3010 blocks contain four numbers? ______

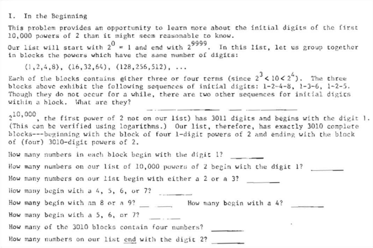
How many numbers on our list end with the digit 2? ______

Plausible answers:10%, 11%

Interesting Question

Interesting Question

For a nice data set, such as the Fibonacci numbers, what percent of the leading digits are 1?



I. In the Beginning

This problem provides an opportunity to learn more about the initial digits of the first 10,000 powers of 2 than it might seem reasonable to know.

Our list will start with $2^0 = 1$ and end with 2^{9999}. In this list, let us group together in blocks the powers which have the same number of digits:

(1,2,4,8), (16,32,64), (128,256,512), ...

Each of the blocks contains either three or four terms (since $2^3 < 10 < 2^4$). The three blocks above exhibit the following sequences of initial digits: 1-2-4-8, 1-3-6, 1-2-5. Though they do not occur for a while, there are two other sequences for initial digits within a block. What are they? ______ ______

$2^{10,000}$, the first power of 2 not on our list) has 3011 digits and begins with the digit 1. (This can be verified using logarithms.) Our list, therefore, has exactly 3010 complete blocks---beginning with the block of four 1-digit powers of 2 and ending with the block of (four) 3010-digit powers of 2.

How many numbers in each block begin with the digit 1? ______

How many numbers on our list of 10,000 powers of 2 begin with the digit 1? ______

How many numbers on our list begin with either a 2 or a 3? ______

How many begin with a 4, 5, 6, or 7? ______

How many begin with an 8 or a 9? ______ How many begin with a 4? ______

How many begin with a 5, 6, or 7? ______

How many of the 3010 blocks contain four numbers? ______

How many numbers on our list end with the digit 2? ______

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

Summary

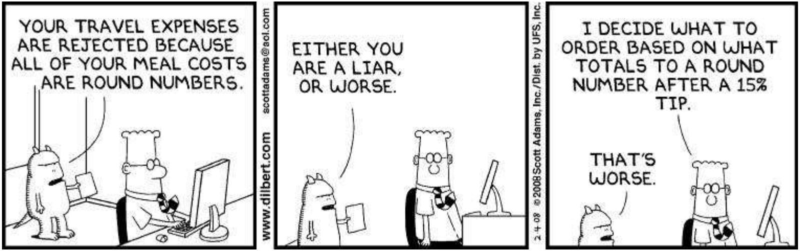
- State Benford's Law.
- Discuss examples and applications.
- Sketch proofs.
- Describe open problems.

Caveats!

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

Caveats!

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



Benford's Law: Newcomb (1881), Benford (1938)

Statement

For many data sets, probability of observing a first digit of d base B is $\log_B\left(\frac{d+1}{d}\right)$; base 10 about 30% are 1s.

Benford's Law: Newcomb (1881), Benford (1938)

Statement

For many data sets, probability of observing a first digit of d base B is $\log_B\left(\frac{d+1}{d}\right)$; base 10 about 30% are 1s.

- Not all data sets satisfy Benford's Law.

Benford's Law: Newcomb (1881), Benford (1938)

Statement

For many data sets, probability of observing a first digit of d base B is $\log_B\left(\frac{d+1}{d}\right)$; base 10 about 30% are 1s.

- Not all data sets satisfy Benford's Law.
  - ⋄ Long street $[1, L]$: $L = 199$ versus $L = 999$.

Benford's Law: Newcomb (1881), Benford (1938)

Statement

For many data sets, probability of observing a first digit of d base B is $\log_B\left(\frac{d+1}{d}\right)$; base 10 about 30% are 1s.

- Not all data sets satisfy Benford's Law.
  - ⋄ Long street $[1, L]$: $L = 199$ versus $L = 999$.
  - ⋄ Oscillates between $1/9$ and $5/9$ with first digit 1.

Benford's Law: Newcomb (1881), Benford (1938)

Statement

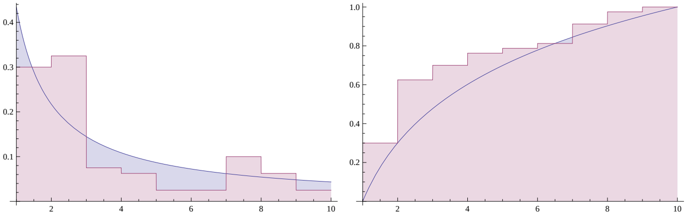
For many data sets, probability of observing a first digit of d base B is $\log_B\left(\frac{d+1}{d}\right)$; base 10 about 30% are 1s.

- Not all data sets satisfy Benford's Law.
  - ◇ Long street $[1, L]$: $L = 199$ versus $L = 999$.
  - ◇ Oscillates between $1/9$ and $5/9$ with first digit 1.
  - ◇ Many streets of different sizes: close to Benford.

Examples

- recurrence relations
- special functions (such as $n!$)
- iterates of power, exponential, rational maps
- products of random variables
- L-functions, characteristic polynomials
- iterates of the $3x + 1$ map
- differences of order statistics
- hydrology and financial data
- many hierarchical Bayesian models

Bastille Day, Wikipedia and Google

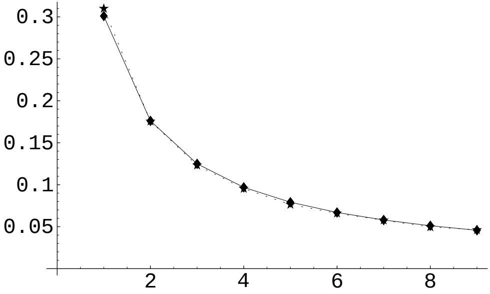


First digit of number of hits Google returns for first 80 distinct words and numbers from Wikipedia's *Bastille Day* entry on July 10, 2011. Chisquare of 46.2, highly non-Benford (99% of time should be at most 20).

Riemann Zeta Function

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

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



Applications

- analyzing round-off errors
- determining the optimal way to store numbers
- detecting tax and image fraud, and data integrity

General Theory

Mantissas

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

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

Key observation: $\log_{10}(x) = \log_{10}(\widetilde{x}) \bmod 1$ if and only if x and $\widetilde{x}$ have the same leading digits. Thus often study $y = \log_{10} x$.

Equidistribution and Benford's Law

Equidistribution

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

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

Equidistribution and Benford's Law

Equidistribution

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

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

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

Equidistribution and Benford's Law

Equidistribution

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

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

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

Equidistribution and Benford's Law

Equidistribution

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

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

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

Equidistribution and Benford's Law

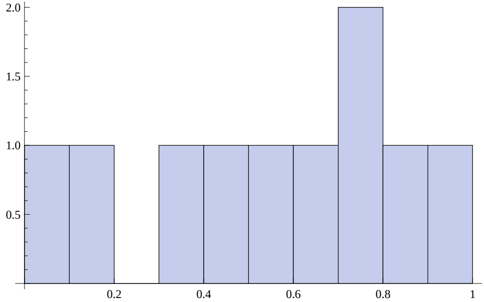
Equidistribution

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

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

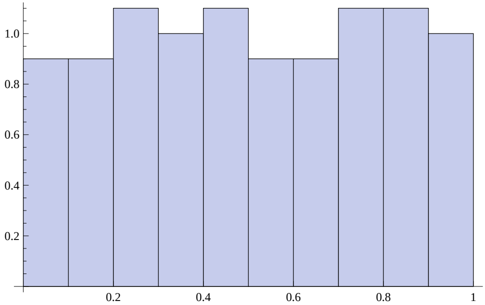
- Thm: $\beta \notin \mathbb{Q}$, $n\beta$ is equidistributed mod 1.
- Examples: $\log_{10} 2, \log_{10}\left(\frac{1+\sqrt{5}}{2}\right) \notin \mathbb{Q}$.
  Proof: if rational: $2 = 10^{p/q}$.
  Thus $2^q = 10^p$ or $2^{q-p} = 5^p$, impossible.

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



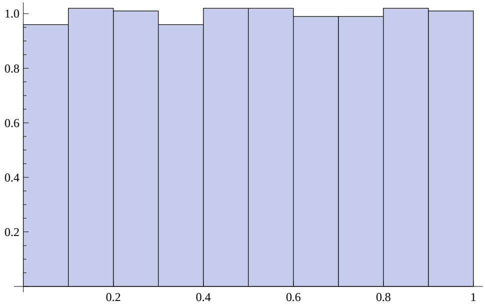
$n\sqrt{\pi}$ mod 1 for $n \leq 10$

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



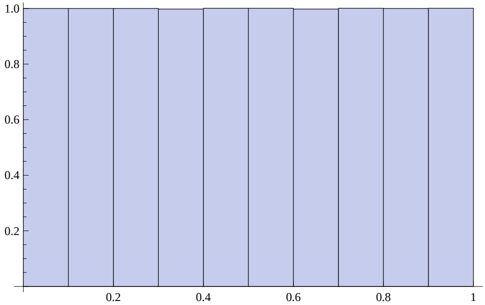
$n\sqrt{\pi}$ mod 1 for $n \leq 100$

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



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

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



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

Logarithms and Benford's Law

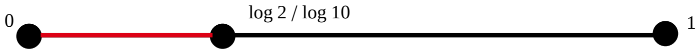
Fundamental Equivalence

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

Logarithms and Benford's Law

Fundamental Equivalence

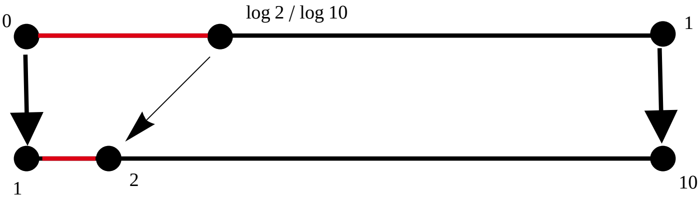
Data set $\{x_i\}$ is Benford base B if $\{y_i\}$ is equidistributed mod 1, where $y_i = \log_B x_i$.



Logarithms and Benford's Law

Fundamental Equivalence

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



Examples

- 2^n is Benford base 10 as $\log_{10} 2 \not\in \mathbb{Q}$.

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				

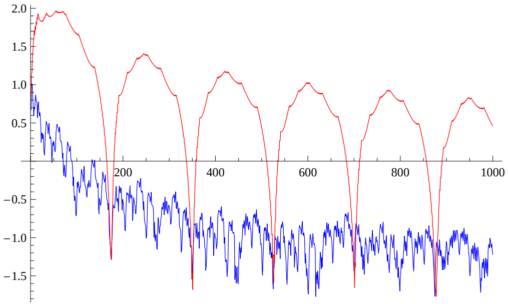
Logarithms and Benford's Law

χ^2 values for α^n, $1 \le n \le N$ (5% 15.5).

N	$\chi^2(\gamma)$	$\chi^2(e)$	$\chi^2(\pi)$
100	0.72	0.30	46.65
200	0.24	0.30	8.58
400	0.14	0.10	10.55
500	0.08	0.07	2.69
700	0.19	0.04	0.05
800	0.04	0.03	6.19
900	0.09	0.09	1.71
1000	0.02	0.06	2.90

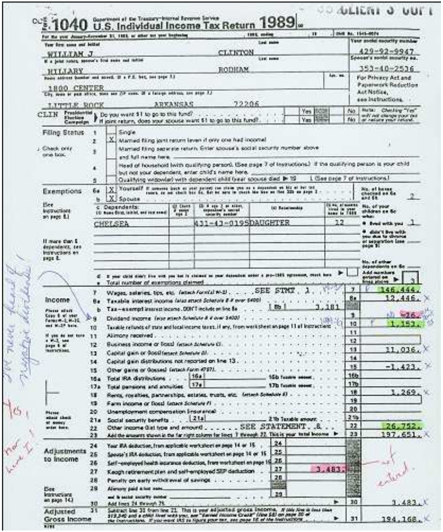
Logarithms and Benford's Law: Base 10

$\log(\chi^2)$ vs N for π^n (red) and e^n (blue), $n \in \{1, \dots, N\}$. Note $\pi^{175} \approx 1.0028 \cdot 10^{87}$, (5%, $\log(\chi^2) \approx 2.74$).



Applications

Applications for the IRS: Detecting Fraud



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

CLIENT'S COPY

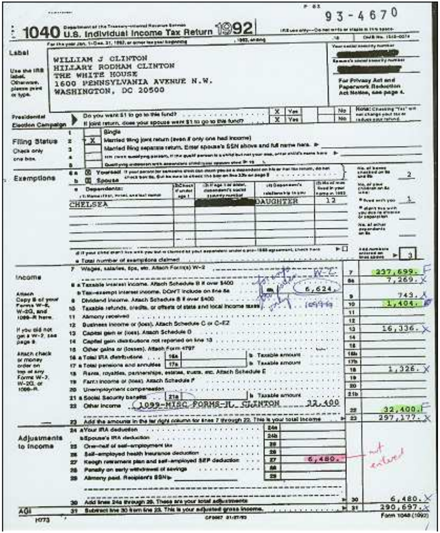
WILLIAM J. CLINTON 429-92-9947
HILLARY RODHAM 353-40-2536
1800 CENTER
LITTLE ROCK ARKANSAS 72206

Filing Status: 2 X Married filing joint return (even if only one had income)

Exemptions: 6a X Yourself; b X Spouse
c Dependents: CHELSEA 431-43-0195 DAUGHTER 12

7	Wages, salaries, tips, etc. SEE STMT 1	146,444.
8a	Taxable interest income	12,446.
8b	Tax-exempt interest income	3,181
9	Dividend income	26.
10	Taxable refunds of state and local income taxes	1,153.
13	Capital gain or (loss)	11,036.
15	Other gains or (losses)	-1,423.
18	Rents, royalties, partnerships, estates, trusts, etc.	1,269.
22	Other income SEE STATEMENT 8	26,752.
23	Add the amounts shown in the far right column for lines 7 through 22. This is your total income	197,651.
27	Keogh retirement plan and self-employed SEP deduction	3,483.
30	Add lines 24 through 29	3,483.
31	Adjusted gross income	194,168.

Applications for the IRS: Detecting Fraud


93-4670

1040 U.S. Individual Income Tax Return 1992

Label

Your social security number [redacted]

Spouse's social security number [redacted]

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

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

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

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

Exemptions — 6a Yourself; b Spouse; c Dependents:

(1) Name	(2) Check if under age 1	(3) Dependent's social security number	(4) Dependent's relationship to you	(5) No. of months lived in your home in 1992
CHELSEA		[redacted]	DAUGHTER	12

2 — 1 — 3

e Total number of exemptions claimed

Income

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

Adjustments to Income

24a	Your IRA deduction	
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. Recipient's SSN	
30	Add lines 24a through 29. These are your total adjustments	6,480.
31	Subtract line 30 from line 23. This is your adjusted gross income.	290,697.

AGI — Form 1040 (1992)

Detecting Fraud

Bank Fraud

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

Detecting Fraud

Bank Fraud

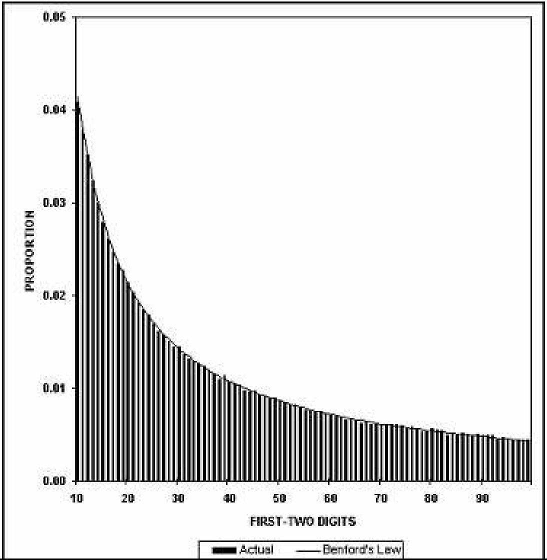
- Audit of a bank revealed huge spike of numbers starting with 48 and 49, most due to one person.

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 protests/complaints over Iran's 2009 elections.

Lot of analysis; data moderately suspicious:

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

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

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

$3x + 1$ Problem

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

$3x + 1$ Problem

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

$3x+1$ Problem

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

$3x + 1$ Problem

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

$3x + 1$ Problem

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

$3x + 1$ Problem

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

$3x+1$ Problem

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

$3x + 1$ Problem

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

Heuristic Proof of $3x+1$ Conjecture

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

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.

$3x+1$ 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

Products of Random Variables

Preliminaries

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

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

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

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

Modulo 1 Central Limit Theorem

Theorem (M– and Nigrini 2007)

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

◇ Gives info on rate of convergence.

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

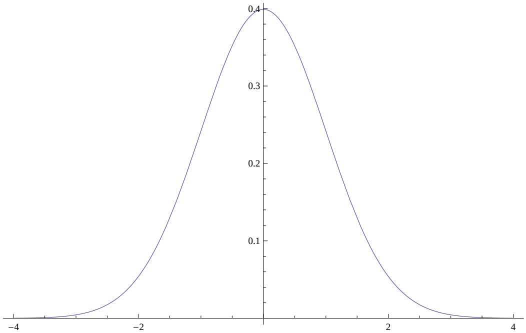


Figure: Plot of normal (mean 0, stdev 1).

Proof under stronger conditions

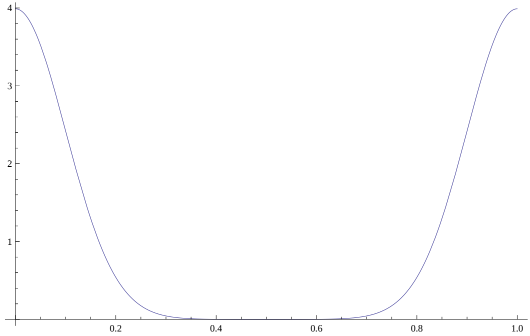


Figure: Plot of normal (mean 0, stdev .1) modulo 1.

Proof under stronger conditions

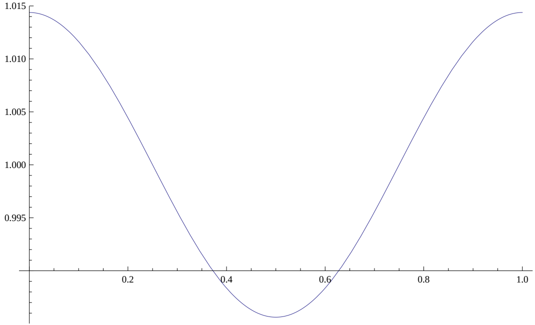


Figure: Plot of normal (mean 0, stdev .5) modulo 1.

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 Journal of Statistics, Series B **30** (1968), 47–58.
- R. N. Bhattacharya, *Speed of convergence of the n-fold convolution of a probability measure ona compact group*, Z. Wahrscheinlichkeitstheorie verw. Geb. **25** (1972), 1–10.
- F. Benford, *The law of anomalous numbers*, Proceedings of the American Philosophical Society **78** (1938), 551–572.
- A. Berger, Leonid A. Bunimovich and T. Hill, *One-dimensional dynamical systems and Benford's Law*, Trans. Amer. Math. Soc. **357** (2005), no. 1, 197–219.

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

R. W. Hamming, *On the distribution of numbers*, Bell Syst. Tech. J. **49** (1970), 1609-1625.

T. Hill, *The first-digit phenomenon*, American Scientist **86** (1996), 358–363.

T. Hill, *A statistical derivation of the significant-digit law*, Statistical Science **10** (1996), 354–363.

P. J. Holewijn, *On the uniform distribuiton of sequences of random variables*, Z. Wahrscheinlichkeitstheorie verw. Geb. **14** (1969), 89–92.

W. Hurlimann, *Benford's Law from 1881 to 2006: a bibliography*, `http://arxiv.org/abs/math/0607168`.

D. Jang, J. U. Kang, A. Kruckman, J. Kudo and S. J. Miller, *Chains of distributions, hierarchical Bayesian models and Benford's Law*, Journal of Algebra, Number Theory: Advances and Applications, volume 1, number 1 (March 2009), 37–60.

- E. Janvresse and T. de la Rue, *From uniform distribution to Benford's law*, Journal of Applied Probability **41** (2004) no. 4, 1203–1210.
- A. Kontorovich and S. J. Miller, *Benford's Law, Values of L-functions and the* $3x+1$ *Problem*, Acta Arith. **120** (2005), 269–297.
- D. Knuth, *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*, Addison-Wesley, third edition, 1997.
- J. Lagarias and K. Soundararajan, *Benford's Law for the* $3x+1$ *Function*, J. London Math. Soc. (2) **74** (2006), no. 2, 289–303.
- S. Lang, *Undergraduate Analysis*, 2nd edition, Springer-Verlag, New York, 1997.

- P. Levy, *L'addition des variables aléatoires définies sur une circonférence*, Bull. de la S. M. F. **67** (1939), 1–41.
- E. Ley, *On the peculiar distribution of the U.S. Stock Indices Digits*, The American Statistician **50** (1996), no. 4, 311–313.
- R. M. Loynes, *Some results in the probabilistic theory of asympototic uniform distributions modulo* 1, Z. Wahrscheinlichkeitstheorie verw. Geb. **26** (1973), 33–41.
- S. J. Miller, *When the Cramér-Rao Inequality provides no information*, Communications in Information and Systems **7** (2007), no. 3, 265–272.
- S. J. Miller and M. Nigrini, *The Modulo* 1 *Central Limit Theorem and Benford's Law for Products*, International Journal of Algebra **2** (2008), no. 3, 119–130.
- S. J. Miller and M. Nigrini, *Order Statistics and Benford's law*, International Journal of Mathematics and Mathematical Sciences, Volume 2008 (2008), Article ID 382948, 19 pages.

- S. J. Miller and R. Takloo-Bighash, *An Invitation to Modern Number Theory*, Princeton University Press, Princeton, NJ, 2006.
- S. Newcomb, *Note on the frequency of use of the different digits in natural numbers*, Amer. J. Math. **4** (1881), 39-40.
- M. Nigrini, *Digital Analysis and the Reduction of Auditor Litigation Risk*. Pages 69–81 in *Proceedings of the 1996 Deloitte & Touche / University of Kansas Symposium on Auditing Problems*, ed. M. Ettredge, University of Kansas, Lawrence, KS, 1996.
- M. Nigrini, *The Use of Benford's Law as an Aid in Analytical Procedures*, Auditing: A Journal of Practice & Theory, **16** (1997), no. 2, 52–67.
- M. Nigrini and S. J. Miller, *Benford's Law applied to hydrology data – results and relevance to other geophysical data*, Mathematical Geology **39** (2007), no. 5, 469–490.

M. Nigrini and S. J. Miller, *Data diagnostics using second order tests of Benford's Law*, Auditing: A Journal of Practice and Theory **28** (2009), no. 2, 305–324.

R. Pinkham, *On the Distribution of First Significant Digits*, The Annals of Mathematical Statistics **32**, no. 4 (1961), 1223-1230.

R. A. Raimi, *The first digit problem*, Amer. Math. Monthly **83** (1976), no. 7, 521–538.

H. Robbins, *On the equidistribution of sums of independent random variables*, Proc. Amer. Math. Soc. **4** (1953), 786–799.

H. Sakamoto, *On the distributions of the product and the quotient of the independent and uniformly distributed random variables*, Tôhoku Math. J. **49** (1943), 243–260.

P. Schatte, *On sums modulo 2π of independent random variables*, Math. Nachr. **110** (1983), 243–261.

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