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

Steven J Miller (Smith / Mount Holyoke / Williams Colleges)

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

Smith College, November 15, 2011

Introduction

Interesting Question

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

Plausible answers:

Introduction

Interesting Question

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

Plausible answers: 10%

Introduction

Interesting Question

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

Plausible answers: 10%, 11%

Introduction

Interesting Question

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

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

Summary

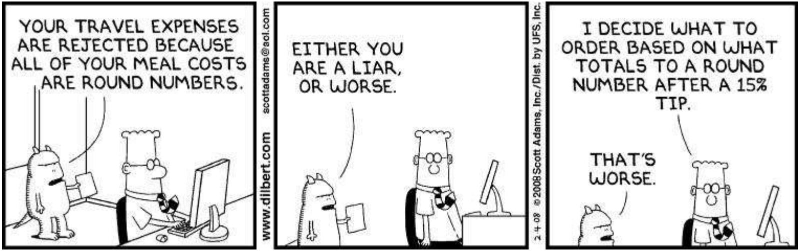
- State Benford's Law.
- Discuss examples and applications.
- Sketch proofs.
- Describe open problems.

Caveats!

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

Caveats!

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


YOUR TRAVEL EXPENSES ARE REJECTED BECAUSE ALL OF YOUR MEAL COSTS ARE ROUND NUMBERS.
scottadams@aol.com
www.dilbert.com
EITHER YOU ARE A LIAR, OR WORSE.
2-4-08 © 2008 Scott Adams, Inc./Dist. by UFS, Inc.
I DECIDE WHAT TO ORDER BASED ON WHAT TOTALS TO A ROUND NUMBER AFTER A 15% TIP.
THAT'S WORSE.

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.
  $\diamond$ 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.
  - $\diamond$ Long street $[1, L]$: $L = 199$ versus $L = 999$.
  - $\diamond$ 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

Applications

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

General Theory

Mantissas (or Significands)

$x \bmod 1$ means the fractional part of x: $x - \lfloor x \rfloor$.

Mantissas (or Significands)

x mod 1 means the fractional part of x: $x - \lfloor x \rfloor$.

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

Mantissas (or Significands)

x mod 1 means the fractional part of x: $x - \lfloor x \rfloor$.

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.

Mantissas (or Significands)

$x \bmod 1$ means the fractional part of x: $x - \lfloor x \rfloor$.

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 $b-a$:

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

- 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 $b-a$:

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

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

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

  Proof: if rational: $2 = 10^{p/q}$.

Equidistribution and Benford's Law

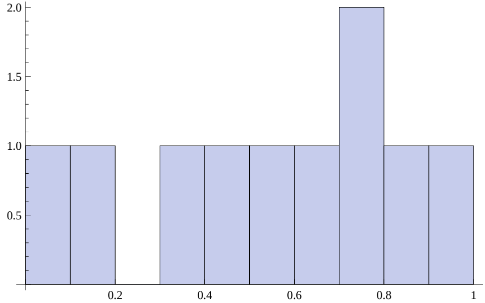
Equidistribution

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

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

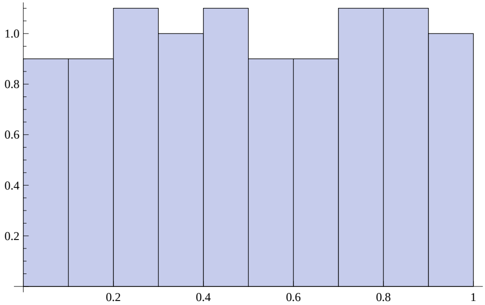
- Thm: $\beta \notin \mathbb{Q}$, $n\beta$ is equidistributed mod 1.
- Examples: $\log_{10} 2, \log_{10}\left(\frac{1+\sqrt{5}}{2}\right) \notin \mathbb{Q}$.
  Proof: if rational: $2 = 10^{p/q}$.
  Thus $2^q = 10^p$ or $2^{q-p} = 5^p$, impossible.

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



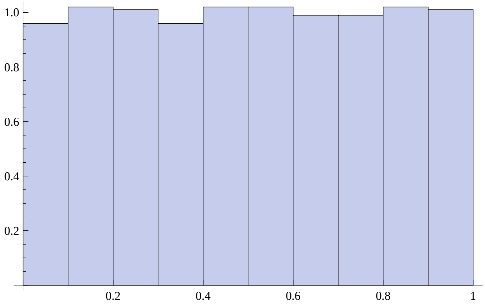
$n\sqrt{\pi} \bmod 1$ for $n \leq 10$

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



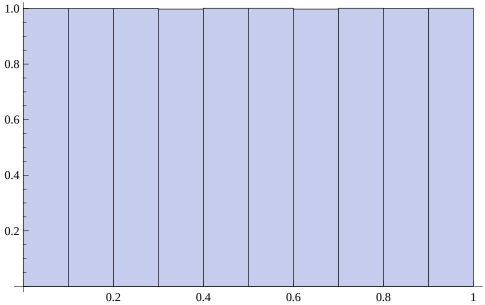
$n\sqrt{\pi}$ mod 1 for $n \leq 100$

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



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

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



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

Logarithms and Benford's Law

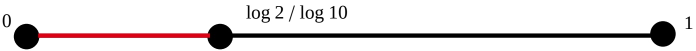
Fundamental Equivalence

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

Logarithms and Benford's Law

Fundamental Equivalence

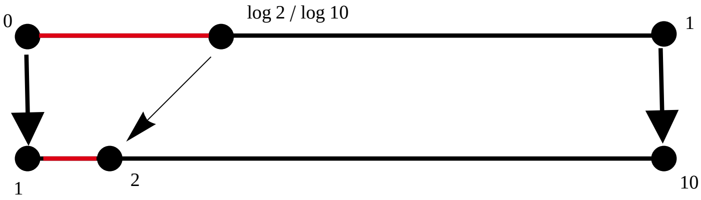
Data set $\{x_i\}$ is Benford base B if $\{y_i\}$ is equidistributed mod 1, where $y_i = \log_B x_i$.



Logarithms and Benford's Law

Fundamental Equivalence

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



Examples

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

Examples

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

- Fibonacci numbers are Benford base 10.

Examples

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

- Fibonacci numbers are Benford base 10.
  $a_{n+1} = a_n + a_{n-1}$.

Examples

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

- Fibonacci numbers are Benford base 10.

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

Examples

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

- Fibonacci numbers are Benford base 10.
  $a_{n+1} = a_n + a_{n-1}$.
  Guess $a_n = r^n$: $r^{n+1} = r^n + r^{n-1}$ or $r^2 = r + 1$.
  Roots $r = (1 \pm \sqrt{5})/2$.

Examples

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

- Fibonacci numbers are Benford base 10.

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

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

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

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

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

- Fibonacci numbers are Benford base 10.

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

Digits of 2^n

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

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

Digits of 2^n

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

1	1024	1048576
2	2048	2097152
4	4096	4194304
8	8192	8388608
16	16384	16777216
32	32768	33554432
64	65536	67108864
128	131072	134217728
256	262144	268435456
512	524288	536870912

digit	#	Obs Prob	Benf Prob
1	18	.300	.301
2	12	.200	.176
3	6	.100	.125
4	6	.100	.097
5	6	.100	.079
6	4	.067	.067
7	2	.033	.058
8	5	.083	.051
9	1	.017	.046

Digits of 2^n

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

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

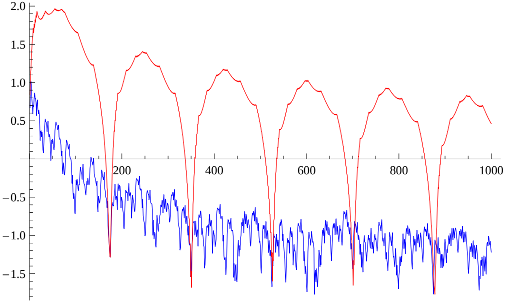
Logarithms and Benford's Law

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

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

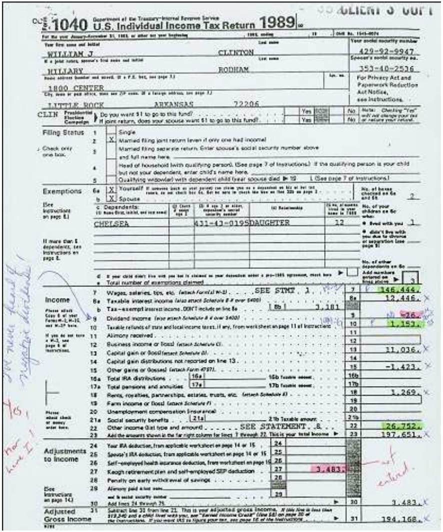
Logarithms and Benford's Law: Base 10

$\log_{10}(\chi^2)$ vs N for π^n (red) and e^n (blue), $n \in \{1, \dots, N\}$. Note $\pi^{175} \approx 1.0028 \cdot 10^{87}$, (5% and 8 d.f., $\log_{10}(\chi^2) \approx .44$).



Applications

Applications for the IRS: Detecting Fraud



CLIENT'S COPY

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

For the year January–December 31, 1989, or other tax year beginning , 1989, ending , 19 . OMB No. 1545-0074

Your first name and initial WILLIAM J	Last name CLINTON	Your social security number 429-92-9947
If a joint return, spouse's first name and initial HILLARY	Last name RODHAM	Spouse's social security no. 353-40-2536
Home address (number and street). If a P.O. box, see page 7.) 1800 CENTER	Apt. no.	For Privacy Act and Paperwork Reduction Act Notice, see Instructions.
City, town or post office, state and ZIP code. (If a foreign address, see page 7.) LITTLE ROCK ARKANSAS 72206		

CLIN

Presidential Election Campaign ▶ Do you want $1 to go to this fund? . . . Yes □ No □ Note: Checking "Yes" will not change your tax or reduce your refund.
If joint return, does your spouse want $1 to go to this fund? . . . Yes □ No □

Filing Status (Check only one box.)
1 □ Single
2 ☒ Married filing joint return (even if only one had income)
3 □ Married filing separate return. Enter spouse's social security number above and full name here.
4 □ Head of household (with qualifying person). (See page 7 of Instructions.) If the qualifying person is your child but not your dependent, enter child's name here.
5 □ Qualifying widow(er) with dependent child (year spouse died ▶ 19). (See page 7 of Instructions.)

Exemptions (See Instructions on page 8.)
6a ☒ Yourself [illegible]
b ☒ Spouse

No. of boxes checked on 6a and 6b 2

c. Dependents: (1) Name (first, initial, and last name)	(2) Check if under age 2	(3) If age 2 or older, dependent's social security number	(4) Relationship	(5) No. of months lived in your home in 1989
CHELSEA		431-43-0195	DAUGHTER	12

If more than 6 dependents, see Instructions on page 8.

No. of your children on 6c who:
- lived with you 1
- didn't live with you due to divorce or separation (see page 9)

No. of other dependents on 6c

d If your child didn't live with you but is claimed as your dependent under a pre-1985 agreement, check here ▶ □
e Total number of exemptions claimed

Add numbers entered on lines above ▶ 3

Income

Please attach Copy B of your Forms W-2, W-2G, and W-2P here.

If you do not have a W-2, see page 6 of Instructions.

Please attach check or money order here.

7	Wages, salaries, tips, etc. (attach Form(s) W-2) . . . . SEE STMT 1		7	146,444.
8a	Taxable interest income (also attach Schedule B if over $400)		8a	12,446.
b	Tax-exempt interest income. DON'T include on line 8a	8b 3,181		
9	Dividend income (also attach Schedule B if over $400)		9	-26.
10	Taxable refunds of state and local income taxes, if any, from worksheet on page 11 of Instructions		10	1,153.
11	Alimony received		11	
12	Business income or (loss) (attach Schedule C)		12	
13	Capital gain or (loss) (attach Schedule D)		13	11,036.
14	Capital gain distributions not reported on line 13		14	
15	Other gains or (losses) (attach Form 4797)		15	-1,423.
16a	Total IRA distributions . . . 16a	16b Taxable amount	16b	
17a	Total pensions and annuities 17a	17b Taxable amount	17b	
18	Rents, royalties, partnerships, estates, trusts, etc. (attach Schedule E)		18	1,269.
19	Farm income or (loss) (attach Schedule F)		19	
20	Unemployment compensation (insurance)		20	
21a	Social security benefits . . . 21a	21b Taxable amount	21b	
22	Other income (list type and amount) . . . . . . SEE STATEMENT 8 . .		22	26,752.
23	Add the amounts shown in the far right column for lines 7 through 22. This is your total income ▶		23	197,651.

Adjustments to Income (See Instructions on page 14.)

24	Your IRA deduction, from applicable worksheet on page 14 or 15	24		
25	Spouse's IRA deduction, from applicable worksheet on page 14 or 15	25		
26	Self-employed health insurance deduction, from worksheet on page 15	26		
27	Keogh retirement plan and self-employed SEP deduction	27	3,483.	
28	Penalty on early withdrawal of savings	28		
29	Alimony paid a. Recipient's last name and b. social security number	29		
30	Add lines 24 through 29. ▶		30	3,483.

Adjusted Gross Income

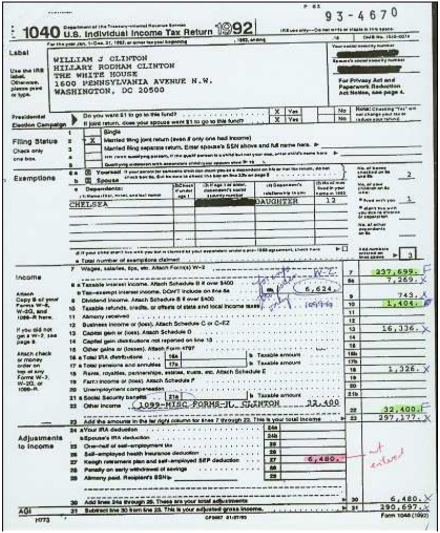
31	Subtract line 30 from line 23. This is your adjusted gross income. If this line is less than $19,340 and a child lived with you, see "Earned Income Credit" (line 58) on page 20 of the Instructions. If you want IRS to figure your tax, see page 16 of the Instructions. ▶	31	194,168.

He never heard of negative dividends

Yes, how I

not entered

Applications for the IRS: Detecting Fraud



P 83 93-4670

1040 Department of the Treasury—Internal Revenue Service
U.S. Individual Income Tax Return 1992
IRS use only—Do not write or staple in this space.
For the year Jan. 1–Dec. 31, 1992, or other tax year beginning , 1992, ending , 19 OMB No. 1545-0074

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

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

Your social security number [redacted]
Spouse's social security number [redacted]
For Privacy Act and Paperwork Reduction Act Notice, see page 4.

Presidential Election Campaign
Do you want $1 to go to this fund? X Yes No
If joint return, does your spouse want $1 to go to this fund? X Yes No
Note: Checking "Yes" will not change your tax or reduce your refund.

Filing Status
Check only one box.
1 □ Single
2 ☒ Married filing joint return (even if only one had income)
3 □ Married filing separate return. Enter spouse's SSN above and full name here. ▶
4 □ Head of household (with qualifying person). If the qualifying person is a child but not your dependent, enter child's name here. ▶
5 □ Qualifying widow(er) with dependent child (year spouse died ▶ 19).

Exemptions
6a ☒ Yourself If your parent (or someone else) can claim you as a dependent on his or her tax return, do not check box 6a. But be sure to check the box on line 33b on page 2.
b ☒ Spouse
c Dependents:

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

No. of boxes checked on 6a and 6b 2
No. of your children on 6c who:
• lived with you 1
• didn't live with you due to divorce or separation
No. of other dependents on 6c
d If your child didn't live with you but is claimed as your dependent under a pre-1985 agreement, check here ▶ □
e Total number of exemptions claimed
Add numbers entered on lines above ▶ 3

Income
Attach Copy B of your Forms W-2, W-2G, and 1099-R here.
If you did not get a W-2, see page 8.
Attach check or money order on top of any Forms W-2, W-2G, or 1099-R.

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

for info only

Adjustments to Income

24a	Your IRA deduction	24a			
b	Spouse's IRA deduction	24b			
25	One-half of self-employment tax	25			
26	Self-employed health insurance deduction	26			
27	Keogh retirement plan and self-employed SEP deduction	27	6,480.		
28	Penalty on early withdrawal of savings	28			
29	Alimony paid. Recipient's SSN ▶	29			
30	Add lines 24a through 29. These are your total adjustments			▶ 30	6,480. X
31	Subtract line 30 from line 23. This is your adjusted gross income.			▶ 31	290,697. X

not entered

AGI

H073 CF0067 01/27/93 Form 1040 (1992)

Detecting Fraud

Bank Fraud

- Audit of a bank revealed huge spike of numbers

Detecting Fraud

Bank Fraud

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

Detecting Fraud

Bank Fraud

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

Detecting Fraud

Bank Fraud

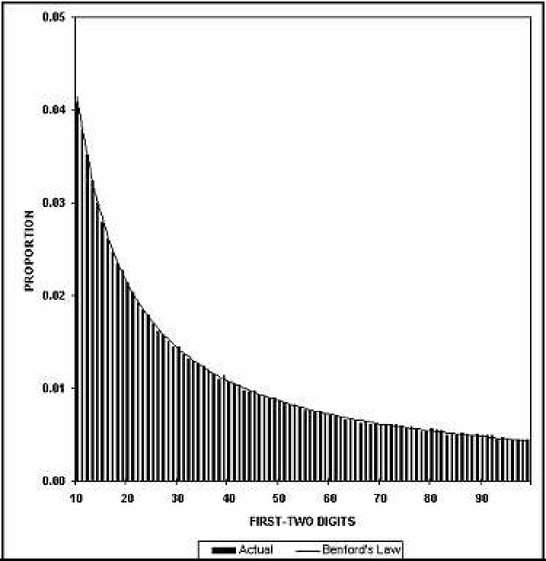
- Audit of a bank revealed huge spike of numbers starting with 48 and 49, most due to one person.

Detecting Fraud

Bank Fraud

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

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 $n = n(x)$, $T^n(x) = 1$.
- $7 \to_1 11 \to_1 17 \to_2 13 \to_3 5 \to_4 1 \to_2 1$

$3x+1$ 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.

Sketch of the proof

- Failed Proof: lattices, bad errors.

- CLT: $(S_m - 2m)/\sqrt{2m} \to N(0,1)$:

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

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

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

Sketch of the proof

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

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

$3x + 1$ 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

Copulas and Benford's Law (joint with Thealexa Becker '13)

Definition of Copulas

Copula: A form of joint CDF between multiple variables with given uniform marginals on the d-dimensional unit cube.

Sklar’s Theorem

Let X and Y be random variables with joint distribution function H and marginal distribution fucntions F and G respectively. There exists a copula, C, such that

$$\text{for all } x, y \in \mathbb{R}, \quad H(x, y) = C(F(x), G(y)).$$

Archimedean Copulas

A commonly used / studied family of copulas is of the form

$$C(x,y) \;=\; \phi^{-1}(\phi(x)+\phi(y))$$

where ϕ is the generator and ϕ^{-1} is the inverse generator of the copula.

Investigating the Benfordness of the product of random variables arising from copulas.

Clayton Copula: $C(x,y) = (x^{-\theta} + y^{-\theta} - 1)^{-1/\theta}$.

PDF (bivariate): $\theta(\theta^{-1}+1)(xy)^{-\theta-1}(x^{-\theta}+y^{-\theta}-1)^{-2-1/\theta}$.

PDF (general case):

$\theta^{n-1}\frac{\Gamma(n+\theta^{-1})}{\Gamma(1+\theta^{-1})}(x_1\cdots x_n)^{-\theta-1}(x_1^{-\theta}+\cdots+x_n^{-\theta}-1)^{-n-1/\theta}$.

Results

- Early data and chi-square tests of multivariate copulas suggest Benford behavior of the products of copulas.
- Proof strategy includes the integration of the PDF over the region in which the product has first digit *d* using Poisson summation:

$$\int_0^1 \cdots \int_0^1 \sum_k \widehat{\phi}_{\log_{10}(x_1 \cdots x_n)}(k) p(x_1, \ldots, x_n) dx_1 \cdots dx_n,$$

where

$$\phi_a(u) = \chi_{[1,2)}(10^{u+a}) = \begin{cases} 1 & \text{if } 10^{u+a} \in [1,2) \\ 0 & \text{otherwise.} \end{cases}$$

Conclusions

Conclusions and Future Investigations

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