

# 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/sjmillier/`

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

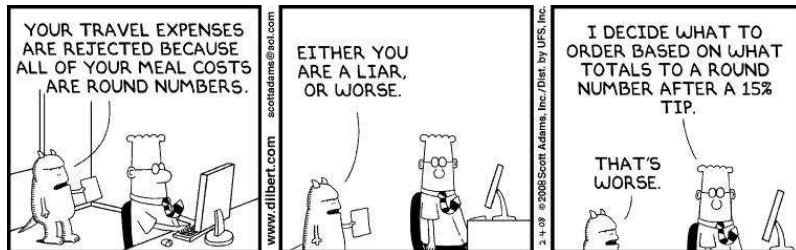
- State Benford's Law.
- Discuss examples and applications.
- Sketch proofs.
- Describe open problems.

## 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.



## 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

## 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)

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

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

## Mantissas (or Significands)

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

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

$M_{10}(x) = M_{10}(\tilde{x})$  if and only if  $x$  and  $\tilde{x}$  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.

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

**Key observation:**  $\log_{10}(x) = \log_{10}(\tilde{x}) \bmod 1$  if and only if  $x$  and  $\tilde{x}$  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  $b - a$ :

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

## Equidistribution and Benford's Law

### Equidistribution

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

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

- 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  $b - a$ :

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

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

## 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

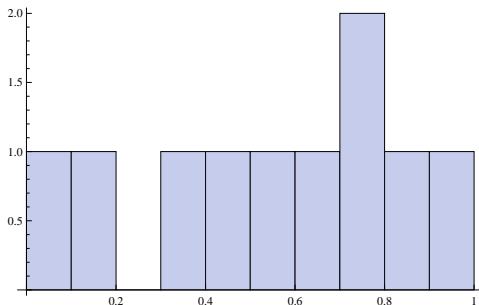
### Equidistribution

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

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

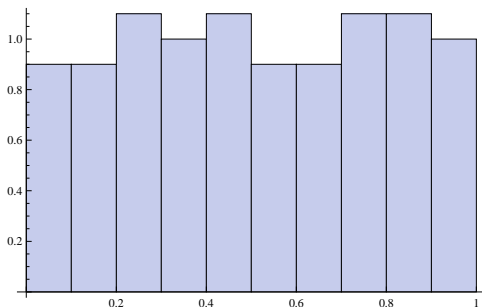
- Thm:  $\beta \notin \mathbb{Q}$ ,  $n\beta$  is equidistributed mod 1.
- Examples:  $\log_{10} 2, \log_{10} \left(\frac{1+\sqrt{5}}{2}\right) \notin \mathbb{Q}$ .  
*Proof:* if rational:  $2 = 10^{p/q}$ .  
 Thus  $2^q = 10^p$  or  $2^{q-p} = 5^p$ , impossible.

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



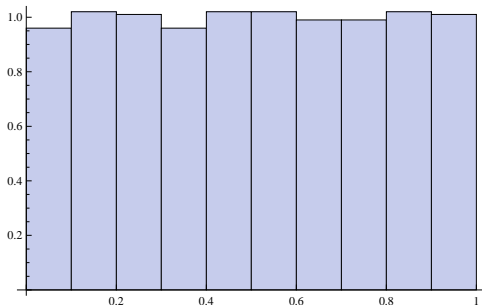
$n\sqrt{\pi} \bmod 1$  for  $n \leq 10$

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



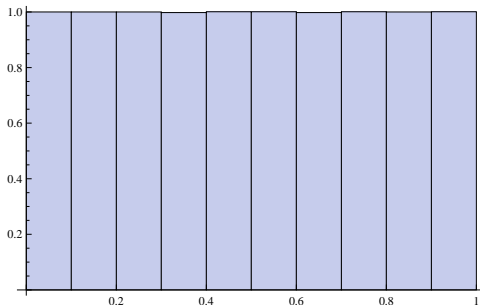
$n\sqrt{\pi} \bmod 1$  for  $n \leq 100$

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



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

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



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

## Logarithms and Benford's Law

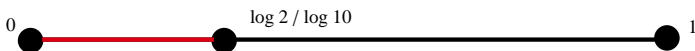
### Fundamental Equivalence

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

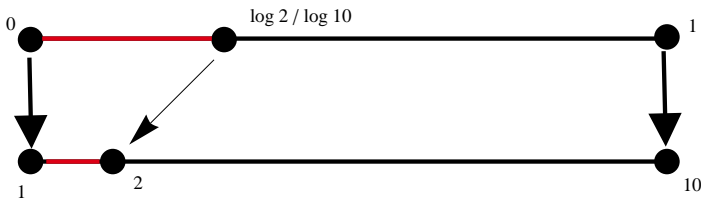
Data set  $\{x_i\}$  is Benford base  $B$  if  $\{y_i\}$  is equidistributed mod 1, where  $y_i = \log_B x_i$ .



## 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  $\log_{10} 2 \notin \mathbb{Q}$ .
- 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.

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

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

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

$$\text{General solution: } a_n = c_1 r_1^n + c_2 r_2^n.$$

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

## Examples

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

- Fibonacci numbers are Benford base 10.

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

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

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

$$\text{General solution: } a_n = c_1 r_1^n + c_2 r_2^n.$$

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

- Most linear recurrence relations Benford.

# Digits of $2^n$

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

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

# Digits of $2^n$

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

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

# Digits of $2^n$

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

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

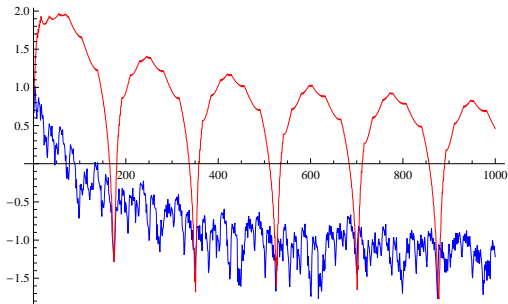
## Logarithms and Benford's Law

$\chi^2$  values for  $\alpha^n$ ,  $1 \leq n \leq N$  (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  $\pi^n$  (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

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

For the year **1989** or other tax year beginning **1989**, ending **1989** (OMB No. 1545-0047)

**WILLIAM J. CLINTON** Last name  
**HILJARY BODHAM** Last name  
**1800 CENTER** Street or rural route, box or P.O. box, see page 11  
**ARKANSAS 72206** City or town, state, and ZIP code, see page 11

**CLIN** Presidential Election Campaign  Do you want \$1 to go to this fund?  Yes  No  Do you want \$1 to go to this fund?  Yes  No  Do you want \$1 to go to this fund?  Yes  No  Do you want \$1 to go to this fund?  Yes  No  Do you want \$1 to go to this fund?  Yes  No

Filing Status **1** Single  **2** Married filing joint return (even if only one had income)  **3** Married filing separate returns. Enter spouse's social security number above and full name here.  **4** Head of household (must be 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 (your spouse died in 1981). (See page 7 of instructions.)

Exemptions **6a**  Yourself  Spouse  Dependent  Other  Total number of exemptions claimed **7** **3**

Income **7** Wages, salaries, tips, etc. (attach Form(s)) **84** **12,446.**  
**8a** Taxable interest income (attach Schedule B if over \$400) **1** **1,381**  
**9** Dividend income (attach Schedule D over \$400) **3** **1,926.**  
**10** Taxable refunds of state and local income taxes, if any, from worksheet on page 11 of instructions **10** **11,153.**  
**11** Alimony received **11** **0.**  
**12** Business income or loss (attach Schedule C) **12** **0.**  
**13** Capital gain or loss (attach Schedule D) **14** **31,036.**  
**14** Capital gain distributions not reported on line 13 **15** **0.**  
**15** Other gains or losses (attach Form 970) **16** **0.**  
**16a** Total IRA distributions **16a** **16b** **16c** **16d** **16e** **16f** **16g** **16h** **16i** **16j** **16k** **16l** **16m** **16n** **16o** **16p** **16q** **16r** **16s** **16t** **16u** **16v** **16w** **16x** **16y** **16z** **16aa** **16ab** **16ac** **16ad** **16ae** **16af** **16ag** **16ah** **16ai** **16aj** **16ak** **16al** **16am** **16an** **16ao** **16ap** **16aq** **16ar** **16as** **16at** **16au** **16av** **16aw** **16ax** **16ay** **16az** **16ba** **16bb** **16bc** **16bd** **16be** **16bf** **16bg** **16bh** **16bi** **16bj** **16bk** **16bl** **16bm** **16bn** **16bo** **16bp** **16bq** **16br** **16bs** **16bt** **16bu** **16bv** **16bw** **16bx** **16by** **16bz** **16ca** **16cb** **16cc** **16cd** **16ce** **16cf** **16cg** **16ch** **16ci** **16cj** **16ck** **16cl** **16cm** **16cn** **16co** **16cp** **16cq** **16cr** **16cs** **16ct** **16cu** **16cv** **16cw** **16cx** **16cy** **16cz** **16da** **16db** **16dc** **16dd** **16de** **16df** **16dg** **16dh** **16di** **16dj** **16dk** **16dl** **16dm** **16dn** **16do** **16dp** **16dq** **16dr** **16ds** **16dt** **16du** **16dv** **16dw** **16dx** **16dy** **16dz** **16ea** **16eb** **16ec** **16ed** **16ee** **16ef** **16eg** **16eh** **16ei** **16ej** **16ek** **16el** **16em** **16en** **16eo** **16ep** **16eq** **16er** **16es** **16et** **16eu** **16ev** **16ew** **16ex** **16ey** **16ez** **16fa** **16fb** **16fc** **16fd** **16fe** **16ff** **16fg** **16fh** **16fi** **16fj** **16fk** **16fl** **16fm** **16fn** **16fo** **16fp** **16fq** **16fr** **16fs** **16ft** **16fu** **16fv** **16fw** **16fx** **16fy** **16fz** **16ga** **16gb** **16gc** **16gd** **16ge** **16gf** **16gg** **16gh** **16gi** **16gj** **16gk** **16gl** **16gm** **16gn** **16go** **16gp** **16gq** **16gr** **16gs** **16gt** **16gu** **16gv** **16gw** **16gx** **16gy** **16gz** **16ha** **16hb** **16hc** **16hd** **16he** **16hf** **16hg** **16hh** **16hi** **16hj** **16hk** **16hl** **16hm** **16hn** **16ho** **16hp** **16hq** **16hr** **16hs** **16ht** **16hu** **16hv** **16hw** **16hx** **16hy** **16hz** **16ia** **16ib** **16ic** **16id** **16ie** **16if** **16ig** **16ih** **16ii** **16ij** **16ik** **16il** **16im** **16in** **16io** **16ip** **16iq** **16ir** **16is** **16it** **16iu** **16iv** **16iw** **16ix** **16iy** **16iz** **16ja** **16jb** **16jc** **16jd** **16je** **16jf** **16jg** **16jh** **16ji** **16jj** **16jk** **16jl** **16jm** **16jn** **16jo** **16jp** **16jq** **16jr** **16js** **16jt** **16ju** **16jv** **16jw** **16jx** **16jy** **16jz** **16ka** **16kb** **16kc** **16kd** **16ke** **16kf** **16kg** **16kh** **16ki** **16kj** **16kl** **16km** **16kn** **16ko** **16kp** **16kq** **16kr** **16ks** **16kt** **16ku** **16kv** **16kw** **16kx** **16ky** **16kz** **16la** **16lb** **16lc** **16ld** **16le** **16lf** **16lg** **16lh** **16li** **16lj** **16lk** **16ll** **16lm** **16ln** **16lo** **16lp** **16lq** **16lr** **16ls** **16lt** **16lu** **16lv** **16lw** **16lx** **16ly** **16lz** **16ma** **16mb** **16mc** **16md** **16me** **16mf** **16mg** **16mh** **16mi** **16mj** **16mk** **16ml** **16mn** **16mo** **16mp** **16mq** **16mr** **16ms** **16mt** **16mu** **16mv** **16mw** **16mx** **16my** **16mz** **16na** **16nb** **16nc** **16nd** **16ne** **16nf** **16ng** **16nh** **16ni** **16nj** **16nk** **16nl** **16nm** **16nn** **16no** **16np** **16nq** **16nr** **16ns** **16nt** **16nu** **16nv** **16nw** **16nx** **16ny** **16nz** **16oa** **16ob** **16oc** **16od** **16oe** **16of** **16og** **16oh** **16oi** **16oj** **16ok** **16ol** **16om** **16on** **16oo** **16op** **16oq** **16or** **16os** **16ot** **16ou** **16ov** **16ow** **16ox** **16oy** **16oz** **16pa** **16pb** **16pc** **16pd** **16pe** **16pf** **16pg** **16ph** **16pi** **16pj** **16pk** **16pl** **16pm** **16pn** **16po** **16pp** **16pq** **16pr** **16ps** **16pt** **16pu** **16pv** **16pw** **16px** **16py** **16pz** **16qa** **16qb** **16qc** **16qd** **16qe** **16qf** **16qg** **16qh** **16qi** **16qj** **16qk** **16ql** **16qm** **16qn** **16qo** **16qp** **16qq** **16qr** **16qs** **16qt** **16qu** **16qv** **16qw** **16qx** **16qy** **16qz** **16ra** **16rb** **16rc** **16rd** **16re** **16rf** **16rg** **16rh** **16ri** **16rj** **16rk** **16rl** **16rm** **16rn** **16ro** **16rp** **16rq** **16rr** **16rs** **16rt** **16ru** **16rv** **16rw** **16rx** **16ry** **16rz** **16sa** **16sb** **16sc** **16sd** **16se** **16sf** **16sg** **16sh** **16si** **16sj** **16sk** **16sl** **16sm** **16sn** **16so** **16sp** **16sq** **16sr** **16ss** **16st** **16su** **16sv** **16sw** **16sx** **16sy** **16sz** **16ta** **16tb** **16tc** **16td** **16te** **16tf** **16tg** **16th** **16ti** **16tj** **16tk** **16tl** **16tm** **16tn** **16to** **16tp** **16tq** **16tr** **16ts** **16tt** **16tu** **16tv** **16tw** **16tx** **16ty** **16tz** **16ua** **16ub** **16uc** **16ud** **16ue** **16uf** **16ug** **16uh** **16ui** **16uj** **16uk** **16ul** **16um** **16un** **16uo** **16up** **16uq** **16ur** **16us** **16ut** **16uu** **16uv** **16uw** **16ux** **16uy** **16uz** **16va** **16vb** **16vc** **16vd** **16ve** **16vf** **16vg** **16vh** **16vi** **16vj** **16vk** **16vl** **16vm** **16vn** **16vo** **16vp** **16vq** **16vr** **16vs** **16vt** **16vu** **16vv** **16vw** **16vx** **16vy** **16vz** **16wa** **16wb** **16wc** **16wd** **16we** **16wf** **16wg** **16wh** **16wi** **16wj** **16wk** **16wl** **16wm** **16wn** **16wo** **16wp** **16wq** **16wr** **16ws** **16wt** **16wu** **16wv** **16ww** **16wx** **16wy** **16wz** **16xa** **16xb** **16xc** **16xd** **16xe** **16xf** **16xg** **16xh** **16xi** **16xj** **16xk** **16xl** **16xm** **16xn** **16xo** **16xp** **16xq** **16xr** **16xs** **16xt** **16xu** **16xv** **16xw** **16xx** **16xy** **16xz** **16ya** **16yb** **16yc** **16yd** **16ye** **16yf** **16yg** **16yh** **16yi** **16yj** **16yk** **16yl** **16ym** **16yn** **16yo** **16yp** **16yq** **16yr** **16ys** **16yt** **16yu** **16yv** **16yw** **16yx** **16yy** **16yz** **16za** **16zb** **16zc** **16zd** **16ze** **16zf** **16zg** **16zh** **16zi** **16zj** **16zk** **16zl** **16zm** **16zn** **16zo** **16zp** **16zq** **16zr** **16zs** **16zt** **16zu** **16zv** **16zw** **16zx** **16zy** **16zz**

Adjustments to Income **24** Your IRA deduction, from applicable worksheet on page 14 or 15 **24** **0.**  
**25** Spouse's IRA deduction, from applicable worksheet on page 14 or 15 **25** **0.**  
**26** Self-employed health insurance deduction, from worksheet on page 16 **26** **0.**  
**27** Raffle, reimbursement and self-employed SEP deduction **27** **3,483.**  
**28** Penalty on early withdrawal of savings **28** **0.**  
**29** Alimony paid (see instructions) **29** **0.**

Gross Income **31** **194,168.**

Adjusted Gross Income **31** **194,168.**

Die Instructions on page 14

Adjusted Gross Income **31** **194,168.**

Gross Income **31** **194,168.**

Handwritten notes: "No. 15 - do not have funds - negative bracket", "No. 15 - do not have funds - same as 1", "not entered" (next to line 27).

## Applications for the IRS: Detecting Fraud

93-4670

1040 U.S. Individual Income Tax Return 1992

Department of the Treasury Internal Revenue Service  
118 West City - One West 10th St. - Kansas City, MO 64108

For the year 1992, 1-1-1992, or other tax year beginning 1992, ending

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

Use the IRS label. Otherwise, please print in type.

Do you want \$1 to go to the President's Election Campaign?  Single  Married (file joint return)  If joint return, does your spouse want \$1 to go to the ACP?  Yes  No

Filing Status:  Married (file joint return) (even if only one had income)  
 Married (file separate return. Show spouse's SSN above and full name. If you have a separate return, it will apply to you, not to your spouse, unless you file a joint return.)  
 Single  
 Head of household (see instructions)  
 Qualifying widow(er) with dependent child (see instructions) (attach Form 1-10)

Exemptions:  Yourself  Spouse  Other dependents (see instructions) (attach Form 1-10)

Dependent's name (last, first, and last initial)	SSN	31 Page 1st item dependent's name (last, first, and last initial)	Relationship to you	Qualifies as dependent (see instructions)	File of tax (use in your name in 1992)	File of tax (use in your name in 1992)
CHYLSEA	0811	DAUSHEWER	DAUGHTER	1	2	1

If you filed a return for 1991, was your tax liability reduced by a prior-year agreement, check this box

7 **Total number of exemptions claimed** 3

Income:

Line	Description	Amount
7	Wages, salaries, tips, etc. (Attach Form W-2)	237,659
8	Taxable interest income. Attach Schedule B if over \$400	7,269
9	Tax-exempt interest income. Do not include on this line	
10	Dividend income. Attach Schedule B if over \$400	743
11	Taxable refunds, credits, or offsets of state and local income taxes	1,404
12	Alimony received	
13	Business income or loss. Attach Schedule C or C-EZ	16,336
14	Capital gain or loss. Attach Schedule D	
15	Other gains or losses. Attach Form 4797	
16	Total IRA distributions	
17	Total pensions and annuities	
18	Rents, royalties, partnerships, estates, trusts, etc. Attach Schedule E	1,328
19	Farmland income or loss. Attach Schedule F	
20	Unemployment compensation	
21	Social Security benefits	
22	Other income. (LOAN-INTEREST FORMS-IL, GANSON)	22,400
23	<b>Total income</b>	<b>324,000</b>
24	<b>Add the amounts in the far right column for lines 7 through 23. This is your total income.</b>	<b>297,177</b>

Adjustments to income:

Line	Description	Amount
25	Spouse's IRA deduction	250
26	Overhead of self-employment tax	28
27	Self-employed health insurance deduction	28
28	Keogh retirement plan and self-employed SEP deduction	6,480
29	Penalty on early withdrawal of savings	28
30	Alimony paid. Attach spouse's SSN	28
31	<b>Total adjustments</b>	<b>6,480</b>
32	<b>Subtract line 31 from line 23. This is your adjusted gross income.</b>	<b>290,697</b>

AGI 290,697 From 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

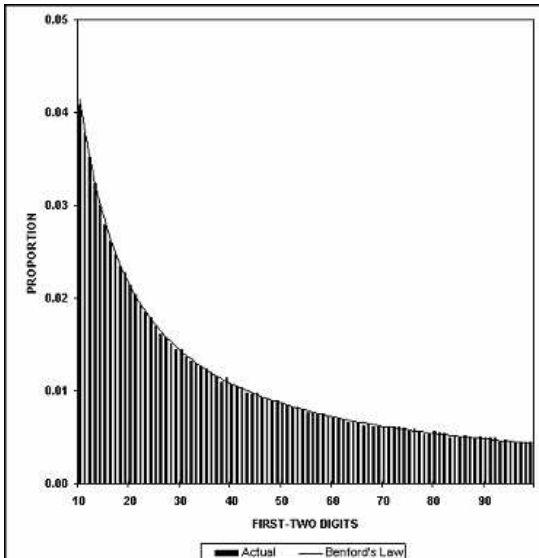
- Audit of a bank revealed huge spike of numbers starting with 48 and 49, most due to one person.

## 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), Erdős (not ready).
- $x$  odd,  $T(x) = \frac{3x+1}{2^k}$ ,  $2^k \parallel 3x + 1$ .
- Conjecture: for some  $n = n(x)$ ,  $T^n(x) = 1$ .
- $7 \rightarrow_1 11 \rightarrow_1 17 \rightarrow_2 13 \rightarrow_3 5 \rightarrow_4 1 \rightarrow_2 1$

## 3x + 1 and Benford

### Theorem (Kontorovich and M–, 2005)

*As  $m \rightarrow \infty$ ,  $x_m / (3/4)^m x_0$  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:  $(S_m - 2m)/\sqrt{2m} \rightarrow N(0, 1)$ :

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

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

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

- Quantified Equidistribution:  $I_\ell = \{\ell M, \dots, (\ell + 1)M - 1\}$ ,  
 $M = m^c$ ,  $c < 1/2$

$$k_1, k_2 \in I_\ell: \left| \eta\left(\frac{k_1}{\sqrt{m}}\right) - \eta\left(\frac{k_2}{\sqrt{m}}\right) \right| \text{ small}$$

$$C = \log_B 2 \text{ of irrationality type } \kappa < 1.2 \cdot 10^{602} < \infty:$$

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

## 3x + 1 Data: random 10,000 digit number, $2^k \parallel 3x + 1$

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 functions  $F$  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  $\phi$  is the generator and  $\phi^{-1}$  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:

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

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.