# Benford's law, or: Why the IRS should care about number theory!

Steven J Miller (Williams College)
Mark Nigrini (The College of New Jersey)

sjm1@williams.edu
http://www.williams.edu/go/math/sjmiller/

Central Connecticut State University
September 25th, 2009

## Summary

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

## Caveats!

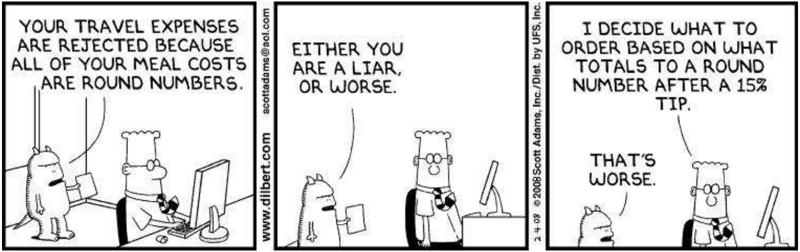
- Not all fraud can be detected by Benford's Law.

## Caveats!

- Not all fraud can be detected by Benford's Law.
- A math test indicating fraud is *not* proof of fraud: unlikely events, alternate reasons.

## Caveats!

- Not all fraud can be detected by Benford's Law.
- A math test indicating fraud is *not* proof of fraud: unlikely events, alternate reasons.



## Notation

- Logarithms: $\log_B x = y$ means $x = B^y$.

## Notation

- Logarithms: $\log_B x = y$ means $x = B^y$.
  - ◇ Example: $\log_{10} 100 = 2$ as $100 = 10^2$.

## Notation

- Logarithms: $\log_B x = y$ means $x = B^y$.
  - ◇ Example: $\log_{10} 100 = 2$ as $100 = 10^2$.
  - ◇ $\log_B(uv) = \log_B u + \log_B v$.

## Notation

- Logarithms: $\log_B x = y$ means $x = B^y$.
  - ⋄ Example: $\log_{10} 100 = 2$ as $100 = 10^2$.
  - ⋄ $\log_B(uv) = \log_B u + \log_B v$.
  - ⋄ $\log_{10}(100 \cdot 1000) = \log_{10}(100) + \log_{10}(1000)$.

## Notation

- Logarithms: $\log_B x = y$ means $x = B^y$.
  - ◇ Example: $\log_{10} 100 = 2$ as $100 = 10^2$.
  - ◇ $\log_B(uv) = \log_B u + \log_B v$.
  - ◇ $\log_{10}(100 \cdot 1000) = \log_{10}(100) + \log_{10}(1000)$.
- Set Theory:
  - ◇ $\mathbb{Q}$ = rational numbers $= \{p/q : p, q \text{ integers}\}$.

## Notation

- Logarithms: $\log_B x = y$ means $x = B^y$.
  - ⋄ Example: $\log_{10} 100 = 2$ as $100 = 10^2$.
  - ⋄ $\log_B(uv) = \log_B u + \log_B v$.
  - ⋄ $\log_{10}(100 \cdot 1000) = \log_{10}(100) + \log_{10}(1000)$.
- Set Theory:
  - ⋄ $\mathbb{Q}$ = rational numbers $= \{p/q : p, q \text{ integers}\}$.
  - ⋄ $x \in S$ means $x$ belongs to $S$.

## Notation

- Logarithms: $\log_B x = y$ means $x = B^y$.
  - ⋄ Example: $\log_{10} 100 = 2$ as $100 = 10^2$.
  - ⋄ $\log_B(uv) = \log_B u + \log_B v$.
  - ⋄ $\log_{10}(100 \cdot 1000) = \log_{10}(100) + \log_{10}(1000)$.
- Set Theory:
  - ⋄ $\mathbb{Q}$ = rational numbers $= \{p/q : p, q \text{ integers}\}$.
  - ⋄ $x \in S$ means $x$ belongs to $S$.
  - ⋄ $[a, b] = \{x : a \leq x \leq b\}$.

## Notation

- Logarithms: $\log_B x = y$ means $x = B^y$.
  - ⋄ Example: $\log_{10} 100 = 2$ as $100 = 10^2$.
  - ⋄ $\log_B(uv) = \log_B u + \log_B v$.
  - ⋄ $\log_{10}(100 \cdot 1000) = \log_{10}(100) + \log_{10}(1000)$.
- Set Theory:
  - ⋄ $\mathbb{Q}$ = rational numbers $= \{p/q : p, q \text{ integers}\}$.
  - ⋄ $x \in S$ means $x$ belongs to $S$.
  - ⋄ $[a, b] = \{x : a \leq x \leq b\}$.
- Modulo 1:

## Notation

- Logarithms: $\log_B x = y$ means $x = B^y$.
  - ⋄ Example: $\log_{10} 100 = 2$ as $100 = 10^2$.
  - ⋄ $\log_B(uv) = \log_B u + \log_B v$.
  - ⋄ $\log_{10}(100 \cdot 1000) = \log_{10}(100) + \log_{10}(1000)$.
- Set Theory:
  - ⋄ $\mathbb{Q}$ = rational numbers $= \{p/q : p, q \text{ integers}\}$.
  - ⋄ $x \in S$ means $x$ belongs to $S$.
  - ⋄ $[a, b] = \{x : a \leq x \leq b\}$.
- Modulo 1:
  - ⋄ Any $x$ can be written as integer + fraction.

## Notation

- Logarithms: $\log_B x = y$ means $x = B^y$.
  - ◇ Example: $\log_{10} 100 = 2$ as $100 = 10^2$.
  - ◇ $\log_B(uv) = \log_B u + \log_B v$.
  - ◇ $\log_{10}(100 \cdot 1000) = \log_{10}(100) + \log_{10}(1000)$.
- Set Theory:
  - ◇ $\mathbb{Q}$ = rational numbers $= \{p/q : p, q \text{ integers}\}$.
  - ◇ $x \in S$ means $x$ belongs to $S$.
  - ◇ $[a, b] = \{x : a \leq x \leq b\}$.
- Modulo 1:
  - ◇ Any $x$ can be written as integer + fraction.
  - ◇ $x$ mod 1 means just the fractional part.

## Notation

- Logarithms: $\log_B x = y$ means $x = B^y$.
  - ⋄ Example: $\log_{10} 100 = 2$ as $100 = 10^2$.
  - ⋄ $\log_B(uv) = \log_B u + \log_B v$.
  - ⋄ $\log_{10}(100 \cdot 1000) = \log_{10}(100) + \log_{10}(1000)$.
- Set Theory:
  - ⋄ $\mathbb{Q}$ = rational numbers = $\{p/q : p, q \text{ integers}\}$.
  - ⋄ $x \in S$ means $x$ belongs to $S$.
  - ⋄ $[a, b] = \{x : a \le x \le b\}$.
- Modulo 1:
  - ⋄ Any $x$ can be written as integer + fraction.
  - ⋄ $x$ mod 1 means just the fractional part.
  - ⋄ Example: $\pi$ mod 1 is about .14159.

## 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

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

## 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.

## 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 $b - a$:

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

## 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 $b - a$:

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

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

## Equidistribution and Benford's Law

### Equidistribution

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

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

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

## Equidistribution and Benford's Law

### Equidistribution

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

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

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

## Equidistribution and Benford's Law

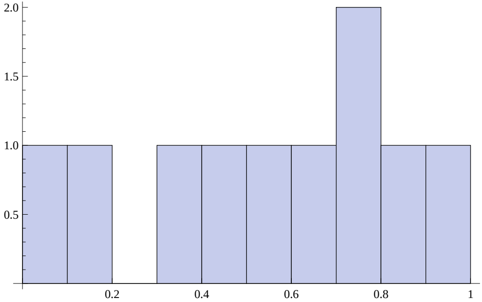
### Equidistribution

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

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

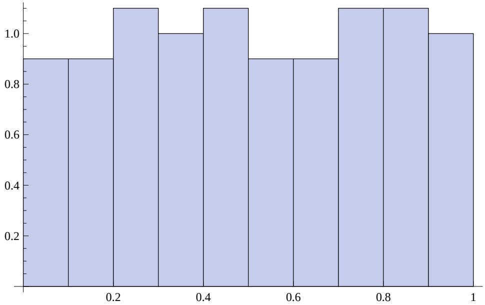
- Thm: $\beta \notin \mathbb{Q}$, $n\beta$ is equidistributed mod 1.
- Examples: $\log_{10} 2, \log_{10}\left(\frac{1+\sqrt{5}}{2}\right) \notin \mathbb{Q}$.
  *Proof:* if rational: $2 = 10^{p/q}$.
  Thus $2^q = 10^p$ or $2^{q-p} = 5^p$, impossible.

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



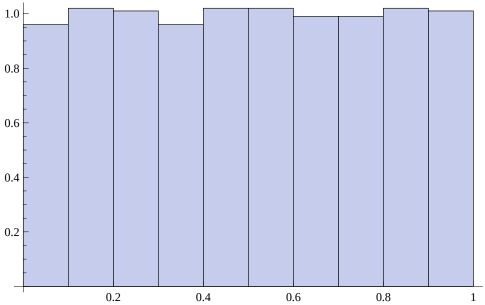

$n\sqrt{\pi}$ mod 1 for $n \leq 10$

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



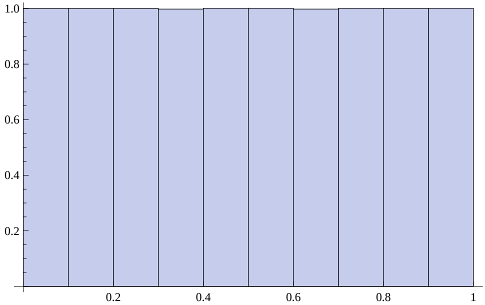

$n\sqrt{\pi}$ mod 1 for $n \leq 100$

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




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

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




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

## Denseness

### Dense

A sequence $\{z_n\}_{n=1}^{\infty}$ of numbers in $[0,1]$ is dense if for any interval $[a,b]$ there are infinitely many $z_n$ in $[a,b]$.

- Dirichlet's Box (or Pigeonhole) Principle:
  If $n+1$ objects are placed in $n$ boxes, at least one box has two objects.

- Denseness of $n\alpha$:
  Thm: If $\alpha \notin \mathbb{Q}$ then $z_n = n\alpha \bmod 1$ is dense.

## Proof $n\alpha$ mod 1 dense if $\alpha \notin \mathbb{Q}$

- Enough to show in $[0, b]$ infinitely often for any $b$.
- Choose any integer $Q > 1/b$.
- $Q$ bins: $\left[0, \frac{1}{Q}\right]$, $\left[\frac{1}{Q}, \frac{2}{Q}\right]$, $\dots$, $\left[\frac{Q-1}{Q}, Q\right]$.
- $Q+1$ objects: $\{\alpha \bmod 1, 2\alpha \bmod 1, \dots, (Q+1)\alpha \bmod 1\}$.
- Two in same bin, say $q_1\alpha$ mod 1 and $q_2\alpha$ mod 1.
- Exists integer $p$ with $0 < q_2\alpha - q_1\alpha - p < \frac{1}{Q}$.
- Get $(q_2 - q_1)\alpha \bmod 1 \in [0, b]$.

## Logarithms and Benford's Law

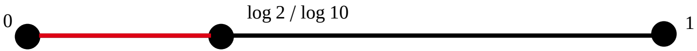
### Fundamental Equivalence

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

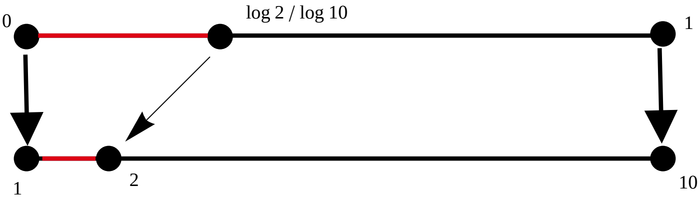
Data set $\{x_i\}$ is Benford base $B$ if $\{y_i\}$ is equidistributed mod 1, where $y_i = \log_B x_i$.



## Logarithms and Benford's Law

### Fundamental Equivalence

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



## Logarithms and Benford's Law

### Fundamental Equivalence

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

**Proof:**

- $x = M_B(x) \cdot B^k$ for some $k \in \mathbb{Z}$.
- $\mathrm{FD}_B(x) = d$ iff $d \leq M_B(x) < d+1$.
- $\log_B d \leq y < \log_B(d+1)$, $y = \log_B x$ mod 1.
- If $Y \sim \mathrm{Unif}(0,1)$ then above probability is $\log_B\left(\frac{d+1}{d}\right)$.

## Examples

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

## Examples

- Fibonacci numbers are Benford base 10.

## Examples

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

## Examples

- Fibonacci numbers are Benford base 10.

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

## Examples

- Fibonacci numbers are Benford base 10.
  $a_{n+1} = a_n + a_{n-1}$.
  Guess $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ 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 $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ 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 $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ 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 $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ 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 $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ 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 $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ 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 $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ 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 | | | | |

## Data Analysis

- $\chi^2$-Tests: Test if theory describes data
  - ◇ Expected probability: $p_d = \log_{10}\left(\frac{d+1}{d}\right)$.
  - ◇ Expect about $Np_d$ will have first digit $d$.
  - ◇ Observe $\mathrm{Obs}(d)$ with first digit $d$.
  - ◇ $\chi^2 = \sum_{d=1}^{9} \frac{(\mathrm{Obs}(d) - Np_d)^2}{Np_d}$.
  - ◇ Smaller $\chi^2$, more likely correct model.

- Will study $\gamma^n$, $e^n$, $\pi^n$.

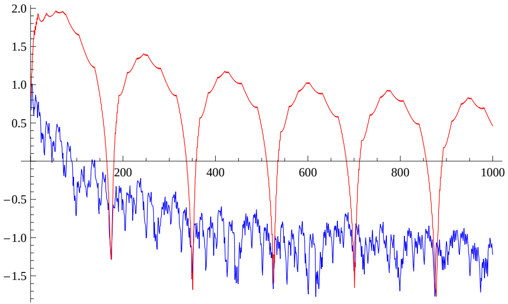
## Logarithms and Benford's Law

$\chi^2$ values for $\alpha^n$, $1 \le n \le N$ (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 |

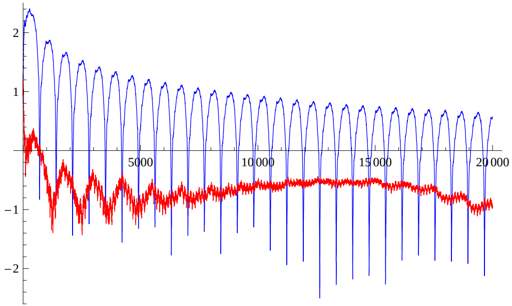
## Logarithms and Benford's Law: Base 10

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



## Logarithms and Benford's Law: Base 20

$\log(\chi^2)$ vs $N$ for $\pi^n$ (red) and $e^n$ (blue), $n \in \{1, \dots, N\}$. Note $e^3 \approx 20.0855$, (5%, $\log(\chi^2) \approx 2.74$).

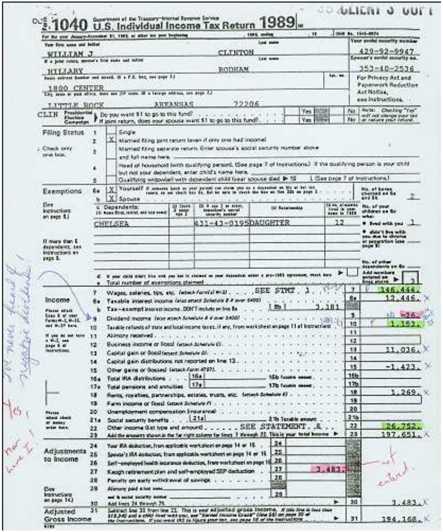


# Applications

## Stock Market

| Milestone | Date | Effective Rate from last milestone |
|---:|---:|---:|
| 108.35 | Jan 12, 1906 | |
| 500.24 | Mar 12, 1956 | 3.0% |
| 1003.16 | Nov 14, 1972 | 4.2% |
| 2002.25 | Jan 8, 1987 | 4.9% |
| 3004.46 | Apr 17, 1991 | 9.5% |
| 4003.33 | Feb 23, 1995 | 7.4% |
| 5023.55 | Nov 21, 1995 | 30.6% |
| 6010.00 | Oct 14, 1996 | 20.0% |
| 7022.44 | Feb 13, 1997 | 46.6% |
| 8038.88 | Jul 16, 1997 | 32.3% |
| 9033.23 | Apr 6, 1998 | 16.1% |
| 10006.78 | Mar 29, 1999 | 10.5% |
| 11209.84 | Jul 16, 1999 | 38.0% |
| 12011.73 | Oct 19, 2006 | 1.0% |
| 13089.89 | Apr 25, 2007 | 16.7% |
| 14000.41 | Jul 19, 2007 | 28.9% |

# Applications for the IRS: Detecting Fraud



CLIENT'S COPY

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

WILLIAM J. CLINTON 429-92-9947

HILLARY RODHAM 353-40-2536

1800 CENTER

LITTLE ROCK ARKANSAS 72206

For Privacy Act and Paperwork Reduction Act Notice, see Instructions.

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

Filling Status
1 Single
2 X 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
6a X Yourself
b X Spouse
c Dependents:
CHELSEA 431-43-0195 DAUGHTER 12

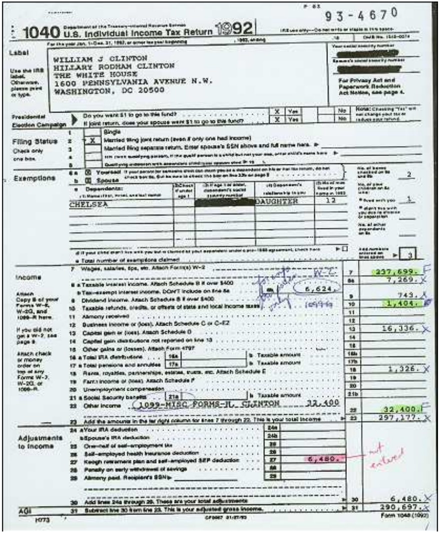
No. of boxes checked on 6a and 6b: 2
No. of your children on 6c who: lived with you: 1
e Total number of exemptions claimed: 3

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

Adjustments to Income
24 Your IRA deduction, from applicable worksheet on page 14 or 15
25 Spouse's IRA deduction, from applicable worksheet on page 14 or 15
26 Self-employed health insurance deduction, from worksheet on page 15
27 Keogh retirement plan and self-employed SEP deduction — 3,483.
28 Penalty on early withdrawal of savings
29 Alimony paid
30 Add lines 24 through 29 — 3,483.

Adjusted Gross Income
31 Subtract line 30 from line 23. This is your adjusted gross income. — 194,168.

# Applications for the IRS: Detecting Fraud


93-4670

1040 Department of the Treasury—Internal Revenue Service
U.S. Individual Income Tax Return 1992

Label

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
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 X Yourself
b X Spouse
c Dependents:
CHELSEA [redacted] DAUGHTER 12

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.
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
17a Total pensions and annuities
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
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
b Spouse's IRA deduction
25 One-half of self-employment tax
26 Self-employed health insurance deduction
27 Keogh retirement plan and self-employed SEP deduction — 6,480.
28 Penalty on early withdrawal of savings
29 Alimony paid.
30 Add lines 24a through 29. These are your total adjustments — 6,480.

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

Form 1040 (1992)

# Applications for the IRS: Detecting Fraud

**Exhibit 3: Check Fraud in Arizona**

The table lists the checks that a manager in the office of the Arizona State Treasurer wrote to divert funds for his own use. The vendors to whom the checks were issued were fictitious.

| Date of Check | Amount |
|---|---|
| October 9, 1992 ↓ | $ 1,927.48 |
| | 27,902.31 |
| October 14, 1992 ↓ | 86,241.90 |
| | 72,117.46 |
| | 81,321.75 |
| | 97,473.96 |
| October 19, 1992 ↓ | 93,249.11 |
| | 89,658.17 |
| | 87,776.89 |
| | 92,105.83 |
| | 79,949.16 |
| | 87,602.93 |
| | 96,879.27 |
| | 91,806.47 |
| | 84,991.67 |
| | 90,831.83 |
| | 93,766.67 |
| | 88,338.72 |
| | 94,639.49 |
| | 83,709.28 |
| | 96,412.21 |
| | 88,432.86 |
| | 71,552.16 |
| TOTAL | $ 1,878,687.58 |

## Applications for the IRS: Detecting Fraud (cont)

- Embezzler started small and then increased dollar amounts.
- Most amounts below $100,000 (critical threshold for data requiring additional scrutiny).
- Over 90% had first digit of 7, 8 or 9.

## Detecting Fraud

### Bank Fraud

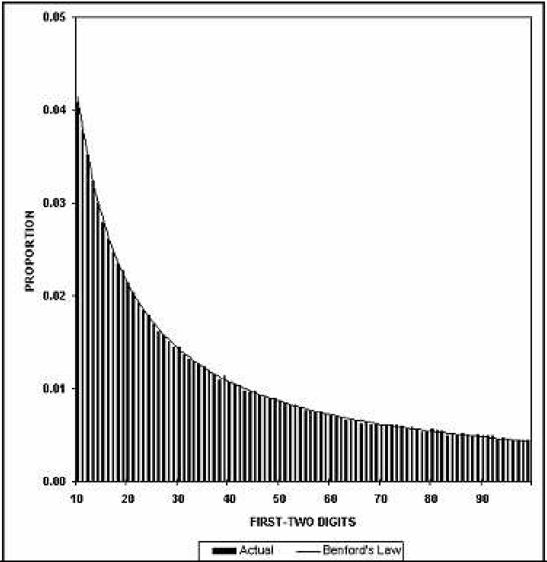
- Audit of a bank revealed huge spike of numbers starting with 48 and 49, most due to one person.
- Write-off limit of $5,000. Officer had friends applying for credit cards, ran up balances just under $5,000 then he would write the debts off.

# Detecting Fraud

## Enron

- Benford's Law detected manipulation of revenue numbers.
- Results showed a tendency towards round Earnings Per Share (0.10, 0.20, etc.).
Consistent with a small but noticeable increase in earnings management in 2002.

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



## Election Fraud: Iran 2009

Numerous protests and complaints over Iran's 2009 elections.
Lot of analysis done; data is moderately suspicious.
Tests done include

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

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

# Benford Good Processes

## Poisson Summation and Benford's Law: Definitions

- Feller, Pinkham (often exact processes)

## Poisson Summation and Benford's Law: Definitions

- Feller, Pinkham (often exact processes)
- data $Y_{T,B} = \log_B \overrightarrow{X}_T$ (discrete/continuous):

$$\mathbb{P}(A) \ = \ \lim_{T\to\infty} \frac{\#\{n \in A : n \le T\}}{T}$$

# Poisson Summation and Benford's Law: Definitions

- Feller, Pinkham (often exact processes)
- data $Y_{T,B} = \log_B \overrightarrow{X}_T$ (discrete/continuous):

$$\mathbb{P}(A) \ = \ \lim_{T\to\infty} \frac{\#\{n \in A : n \le T\}}{T}$$

- Poisson Summation Formula: $f$ nice:

$$\sum_{\ell=-\infty}^{\infty} f(\ell) \ = \ \sum_{\ell=-\infty}^{\infty} \widehat{f}(\ell),$$

$$\text{Fourier transform } \widehat{f}(\xi) \ = \ \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi} dx.$$

## Benford Good Process

$X_T$ is Benford Good if there is a nice $f$ st

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

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

## Benford Good Process

$X_T$ is Benford Good if there is a nice $f$ st

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

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

- Small tails: $G_T(\infty) - G_T(Th(T)) = o(1)$, $G_T(-Th(T)) - G_T(-\infty) = 0(1)$.

## Benford Good Process

$X_T$ is Benford Good if there is a nice $f$ st

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

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

- Small tails: $G_T(\infty) - G_T(Th(T)) = o(1)$, $G_T(-Th(T)) - G_T(-\infty) = 0(1)$.
- Decay of the Fourier Transform: $\sum_{\ell \neq 0} \left| \frac{\widehat{f}(T\ell)}{\ell} \right| = o(1)$.

## Benford Good Process

$X_T$ is Benford Good if there is a nice $f$ st

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

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

- Small tails: $G_T(\infty) - G_T(Th(T)) = o(1)$, $G_T(-Th(T)) - G_T(-\infty) = 0(1)$.
- Decay of the Fourier Transform: $\sum_{\ell \neq 0} \left| \frac{\widehat{f}(T\ell)}{\ell} \right| = o(1)$.
- Small translated error: $\mathcal{E}(a, b, T)) = \sum_{|\ell| \leq Th(T)} [E_T(b + \ell) - E_T(a + \ell)] = o(1)$.

## Main Theorem

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

*$X_T$ converging to $X$ as $T \to \infty$ (think spreading Gaussian). If $X_T$ is Benford good, then $X$ is Benford.*

## Main Theorem

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

*$X_T$ converging to $X$ as $T \to \infty$ (think spreading Gaussian). If $X_T$ is Benford good, then $X$ is Benford.*

- Examples
  - ◇ $L$-functions
  - ◇ characteristic polynomials (RMT)
  - ◇ $3x+1$ problem
  - ◇ geometric Brownian motion.

## Sketch of the proof

- Structure Theorem:
  - ◇ main term is something nice spreading out
  - ◇ apply Poisson summation

## Sketch of the proof

- Structure Theorem:
  - ◇ main term is something nice spreading out
  - ◇ apply Poisson summation

- Control translated errors:
  - ◇ hardest step
  - ◇ techniques problem specific

## Sketch of the proof (continued)

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

## Sketch of the proof (continued)

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

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

## Sketch of the proof (continued)

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

## Sketch of the proof (continued)

$$\begin{aligned}
&\sum_{\ell=-\infty}^{\infty} \mathbb{P}\left(a+\ell \leq \overrightarrow{Y}_{T,B} \leq b+\ell\right) \\
=\;& \sum_{|\ell| \leq Th(T)} \left[G_T(b+\ell) - G_T(a+\ell)\right] + o(1) \\
=\;& \int_a^b \sum_{|\ell| \leq Th(T)} \frac{1}{T} f\left(\frac{t}{T}\right) dt + \mathcal{E}(a,b,T) + o(1) \\
=\;& \widehat{f}(0) \cdot (b-a) + \sum_{\ell \neq 0} \widehat{f}(T\ell) \frac{e^{2\pi i b\ell} - e^{2\pi i a\ell}}{2\pi i\ell} + o(1).
\end{aligned}$$

# Riemann Zeta Function

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

## Riemann Zeta Function

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

$$\begin{aligned}\prod_{p \text{ prime}} \left(1 - \frac{1}{p^s}\right)^{-1} &= \prod_{p \text{ prime}} \left(1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \cdots\right) \\ &= \left(1 + \frac{1}{2^s} + \frac{1}{2^{2s}} + \cdots\right)\left(1 + \frac{1}{3^s} + \frac{1}{3^{2s}} + \cdots\right) \\ &= 1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \frac{1}{5^s} + \frac{1}{(2 \cdot 3)^s} + \cdots\end{aligned}$$

## Riemann Zeta Function

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

$\lim_{s \to 1^+} \zeta(s) = \infty$ implies infinitely many primes.

# Riemann Zeta Function

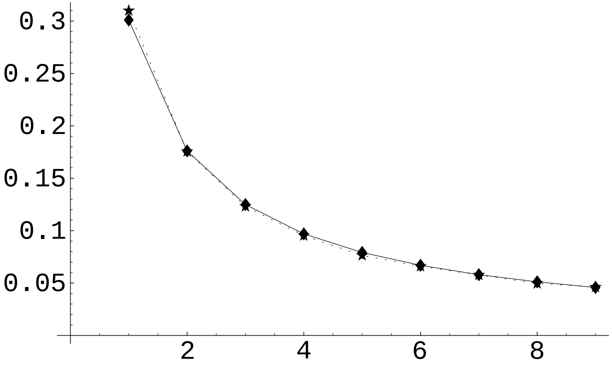
$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{p \text{ prime}} \left(1 - \frac{1}{p^s}\right)^{-1}.$$

$\lim_{s \to 1^+} \zeta(s) = \infty$ implies infinitely many primes.

$\zeta(2) = \pi^2/6$ implies infinitely many primes.

# Riemann Zeta Function

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



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

# $3x + 1$ Problem

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

# $3x + 1$ **Problem**

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

## $3x + 1$ Problem

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

## $3x + 1$ Problem

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

# $3x+1$ Problem

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

# $3x + 1$ Problem

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

## $3x + 1$ 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

$$a_{n+1} = T(a_n)$$

## Heuristic Proof of $3x + 1$ Conjecture

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

# Heuristic Proof of $3x+1$ Conjecture

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

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

# Structure Theorem: Sinai, Kontorovich-Sinai

$$\mathbb{P}(A) = \lim_{N\to\infty} \frac{\#\{n\le N : n\equiv 1,5 \bmod 6, n\in A\}}{\#\{n\le N : n\equiv 1,5 \bmod 6\}}.$$

## Structure Theorem: Sinai, Kontorovich-Sinai

$\mathbb{P}(A) = \lim_{N\to\infty} \frac{\#\{n\le N : n\equiv 1,5 \bmod 6, n\in A\}}{\#\{n\le N : n\equiv 1,5 \bmod 6\}}$.
$(k_1,\dots,k_m)$: two full arithm progressions:
$6\cdot 2^{k_1+\cdots+k_m}p+q$.

## Structure Theorem: Sinai, Kontorovich-Sinai

$\mathbb{P}(A) = \lim_{N\to\infty} \frac{\#\{n\le N: n\equiv 1,5 \bmod 6, n\in A\}}{\#\{n\le N: n\equiv 1,5 \bmod 6\}}$.
$(k_1,\dots,k_m)$: two full arithm progressions:
$6\cdot 2^{k_1+\cdots+k_m}p+q$.

### Theorem (Sinai, Kontorovich-Sinai)

*$k_i$-values are i.i.d.r.v. (geometric, $1/2$):*

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

# Structure Theorem: Sinai, Kontorovich-Sinai

$\mathbb{P}(A) = \lim_{N\to\infty} \frac{\#\{n\le N: n\equiv 1,5 \bmod 6, n\in A\}}{\#\{n\le N: n\equiv 1,5 \bmod 6\}}$.
$(k_1,\dots,k_m)$: two full arithm progressions:
$6\cdot 2^{k_1+\cdots+k_m}p+q$.

## Theorem (Sinai, Kontorovich-Sinai)

*$k_i$-values are i.i.d.r.v. (geometric, $1/2$):*

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

## Structure Theorem: Sinai, Kontorovich-Sinai

$\mathbb{P}(A) = \lim_{N\to\infty} \frac{\#\{n \le N : n \equiv 1,5 \bmod 6, n \in A\}}{\#\{n \le N : n \equiv 1,5 \bmod 6\}}$.
$(k_1, \dots, k_m)$: two full arithm progressions:
$6 \cdot 2^{k_1+\cdots+k_m} p + q$.

### Theorem (Sinai, Kontorovich-Sinai)

*$k_i$-values are i.i.d.r.v. (geometric, $1/2$):*

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

## $3x + 1$ 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.

## Sketch of the proof

- Failed Proof: lattices, bad errors.

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

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

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

## 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

## 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 $\kappa < \infty$:

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

## Irrationality Type

### Irrationality type

$\alpha$ has irrationality type $\kappa$ if $\kappa$ is the supremum of all $\gamma$ with

$$\underline{\lim}_{q\to\infty} q^{\gamma+1} \min_p \left|\alpha - \frac{p}{q}\right| = 0.$$

- Algebraic irrationals: type 1 (Roth's Thm).
- Theory of Linear Forms: $\log_B 2$ of finite type.

## Linear Forms

**Theorem (Baker)**

*$\alpha_1, \dots, \alpha_n$ algebraic numbers height $A_j \geq 4$, $\beta_1, \dots, \beta_n \in \mathbb{Q}$ with height at most $B \geq 4$,*

$$\Lambda = \beta_1 \log \alpha_1 + \cdots + \beta_n \log \alpha_n.$$

*If $\Lambda \neq 0$ then $|\Lambda| > B^{-C\Omega \log \Omega'}$, with $d = [\mathbb{Q}(\alpha_i, \beta_j) : \mathbb{Q}]$, $C = (16nd)^{200n}$, $\Omega = \prod_j \log A_j$, $\Omega' = \Omega / \log A_n$.*

Gives $\log_{10} 2$ of finite type, with $\kappa < 1.2 \cdot 10^{602}$:

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

## Quantified Equidistribution

### Theorem (Erdös-Turan)

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

*There is a C such that for all m:*

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

## Proof of Erdös-Turan

Consider special case $x_n = n\alpha$, $\alpha \notin \mathbb{Q}$.

- Exponential sum $\le \frac{1}{|\sin(\pi h\alpha)|} \le \frac{1}{2||h\alpha||}$.
- Must control $\sum_{h=1}^{m} \frac{1}{h||h\alpha||}$, see irrationality type enter.
- type $\kappa$, $\sum_{h=1}^{m} \frac{1}{h||h\alpha||} = O\left(m^{\kappa-1+\epsilon}\right)$, take $m = \lfloor N^{1/\kappa} \rfloor$.

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

80,514 iterations ($(4/3)^n = a_0$ predicts 80,319); $\chi^2 = 13.5$ (5% 15.5).

| Digit | Number | Observed | Benford |
|---|---|---|---|
| 1 | 24251 | 0.301 | 0.301 |
| 2 | 14156 | 0.176 | 0.176 |
| 3 | 10227 | 0.127 | 0.125 |
| 4 | 7931 | 0.099 | 0.097 |
| 5 | 6359 | 0.079 | 0.079 |
| 6 | 5372 | 0.067 | 0.067 |
| 7 | 4476 | 0.056 | 0.058 |
| 8 | 4092 | 0.051 | 0.051 |
| 9 | 3650 | 0.045 | 0.046 |

## $3x + 1$ Data: random 10,000 digit number, $2|3x + 1$

241,344 iterations, $\chi^2 = 11.4$ (5% 15.5).

| Digit | Number | Observed | Benford |
|---|---|---|---|
| 1 | 72924 | 0.302 | 0.301 |
| 2 | 42357 | 0.176 | 0.176 |
| 3 | 30201 | 0.125 | 0.125 |
| 4 | 23507 | 0.097 | 0.097 |
| 5 | 18928 | 0.078 | 0.079 |
| 6 | 16296 | 0.068 | 0.067 |
| 7 | 13702 | 0.057 | 0.058 |
| 8 | 12356 | 0.051 | 0.051 |
| 9 | 11073 | 0.046 | 0.046 |

# $5x+1$ Data: random 10,000 digit number, $2^k || 5x+1$

27,004 iterations, $\chi^2 = 1.8$ (5% 15.5).

| Digit | Number | Observed | Benford |
|---|---|---|---|
| 1 | 8154 | 0.302 | 0.301 |
| 2 | 4770 | 0.177 | 0.176 |
| 3 | 3405 | 0.126 | 0.125 |
| 4 | 2634 | 0.098 | 0.097 |
| 5 | 2105 | 0.078 | 0.079 |
| 6 | 1787 | 0.066 | 0.067 |
| 7 | 1568 | 0.058 | 0.058 |
| 8 | 1357 | 0.050 | 0.051 |
| 9 | 1224 | 0.045 | 0.046 |

## $5x+1$ Data: random 10,000 digit number, $2|5x+1$

241,344 iterations, $\chi^2 = 3 \cdot 10^{-4}$ (5% 15.5).

| Digit | Number | Observed | Benford |
|---|---|---|---|
| 1 | 72652 | 0.301 | 0.301 |
| 2 | 42499 | 0.176 | 0.176 |
| 3 | 30153 | 0.125 | 0.125 |
| 4 | 23388 | 0.097 | 0.097 |
| 5 | 19110 | 0.079 | 0.079 |
| 6 | 16159 | 0.067 | 0.067 |
| 7 | 13995 | 0.058 | 0.058 |
| 8 | 12345 | 0.051 | 0.051 |
| 9 | 11043 | 0.046 | 0.046 |

# Products
# and
# Chains of Random Variables

## Key Ingredients

- Mellin transform and Fourier transform related by logarithmic change of variable.

- Poisson summation from collapsing to modulo 1 random variables.

## Preliminaries

- $\Xi_1, \dots, \Xi_n$ nice independent r.v.'s on $[0, \infty)$.

## Preliminaries

- $\Xi_1, \dots, \Xi_n$ nice independent r.v.'s on $[0, \infty)$.
- Density $\Xi_1 \cdot \Xi_2$:

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

## Preliminaries

- $\Xi_1, \dots, \Xi_n$ nice independent r.v.'s on $[0, \infty)$.
- Density $\Xi_1 \cdot \Xi_2$:

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

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

$$\begin{aligned} & \int_{t=0}^\infty \mathrm{Prob}\left(\Xi_2 \in \left[0, \frac{x}{t}\right]\right) f_1(t) dt \\ = & \int_{t=0}^\infty F_2\left(\frac{x}{t}\right) f_1(t) dt, \end{aligned}$$

differentiate.

## Mellin Transform

$$\begin{aligned}(\mathcal{M}f)(s) &= \int_0^\infty f(x)x^s \frac{dx}{x} \\ (\mathcal{M}^{-1}g)(x) &= \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} g(s)x^{-s}ds\end{aligned}$$

$$g(s) = (\mathcal{M}f)(s), f(x) = (\mathcal{M}^{-1}g)(x).$$

$$\begin{aligned}(f_1 \star f_2)(x) &= \int_0^\infty f_2\left(\frac{x}{t}\right) f_1(t)\frac{dt}{t} \\ (\mathcal{M}(f_1 \star f_2))(s) &= (\mathcal{M}f_1)(s)\cdot(\mathcal{M}f_2)(s).\end{aligned}$$

# Mellin Transform Formulation: Products Random Variables

## Theorem

*$X_i$'s independent, densities $f_i$. $\Xi_n = X_1 \cdots X_n$,*

$$\begin{aligned} h_n(x_n) &= (f_1 \star \cdots \star f_n)(x_n) \\ (\mathcal{M}h_n)(s) &= \prod_{m=1}^{n} (\mathcal{M}f_m)(s). \end{aligned}$$

*As $n \to \infty$, $\Xi_n$ becomes Benford: $Y_n = \log_B \Xi_n$,*
$|\mathrm{Prob}(Y_n \bmod 1 \in [a,b]) - (b-a)| \leq$

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

# Proof of Kossovsky's Chain Conjecture for certain densities

## Conditions

- $\{\mathcal{D}_i(\theta)\}_{i\in I}$: one-parameter distributions, densities $f_{\mathcal{D}_i(\theta)}$ on $[0,\infty)$.

# Proof of Kossovsky's Chain Conjecture for certain densities

## Conditions

- $\{\mathcal{D}_i(\theta)\}_{i \in I}$: one-parameter distributions, densities $f_{\mathcal{D}_i(\theta)}$ on $[0, \infty)$.
- $p : \mathbb{N} \to I$, $X_1 \sim \mathcal{D}_{p(1)}(1)$, $X_m \sim \mathcal{D}_{p(m)}(X_{m-1})$.

# Proof of Kossovsky's Chain Conjecture for certain densities

## Conditions

- $\{\mathcal{D}_i(\theta)\}_{i\in I}$: one-parameter distributions, densities $f_{\mathcal{D}_i(\theta)}$ on $[0,\infty)$.
- $p:\mathbb{N}\to I$, $X_1 \sim \mathcal{D}_{p(1)}(1)$, $X_m \sim \mathcal{D}_{p(m)}(X_{m-1})$.
- $m \geq 2$,

$$f_m(x_m) \;=\; \int_0^\infty f_{\mathcal{D}_{p(m)}(1)}\left(\frac{x_m}{x_{m-1}}\right) f_{m-1}(x_{m-1})\frac{dx_{m-1}}{x_{m-1}}$$

## Proof of Kossovsky's Chain Conjecture for certain densities

Conditions

- $\{\mathcal{D}_i(\theta)\}_{i\in I}$: one-parameter distributions, densities $f_{\mathcal{D}_i(\theta)}$ on $[0,\infty)$.
- $p : \mathbb{N} \to I$, $X_1 \sim \mathcal{D}_{p(1)}(1)$, $X_m \sim \mathcal{D}_{p(m)}(X_{m-1})$.
- $m \geq 2$,

$$f_m(x_m) \; = \; \int_0^\infty f_{\mathcal{D}_{p(m)}(1)}\left(\frac{x_m}{x_{m-1}}\right) f_{m-1}(x_{m-1}) \frac{dx_{m-1}}{x_{m-1}}$$

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

## Proof of Kossovsky's Chain Conjecture for certain densities

### Theorem (JKKKM)

- *If conditions hold, as $n \to \infty$ the distribution of leading digits of $X_n$ tends to Benford's law.*
- *The error is a nice function of the Mellin transforms: if $Y_n = \log_B X_n$, then*

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

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

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

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

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

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

### Bounds on the error

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

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

# Conclusions

# Conclusions and Future Investigations

- See many different systems exhibit Benford behavior.

## Conclusions and Future Investigations

- See many different systems exhibit Benford behavior.
- Ingredients of proofs (logarithms, equidistribution).

## Conclusions and Future Investigations

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

## Conclusions and Future Investigations

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

# 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*, preprint.

- 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*, to appear in Communications in Information and Systems.

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, *Differences between Independent Variables and Almost Benford Behavior*, preprint. http://arxiv.org/abs/math/0601344

- 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.

- 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 $2\pi$ 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.