Math 341: Probability
Twenty-first Lecture (11/24/09)

Steven J Miller
Williams College

Steven.J.Miller@williams.edu
http://www.williams.edu/go/math/sjmiller/public_html/341/

Bronfman Science Center
Williams College, November 24, 2009

Summary for the Day

Summary for the day

- Benford's Law:
  - Review.
  - Inputs (equidistribution).
  - Clicker question.
  - Difference equations.
  - Products.

- More Sum Than Difference Sets:
  - Definition.
  - Inputs (Chebyshev's Theorem).

Introduction

Caveats!

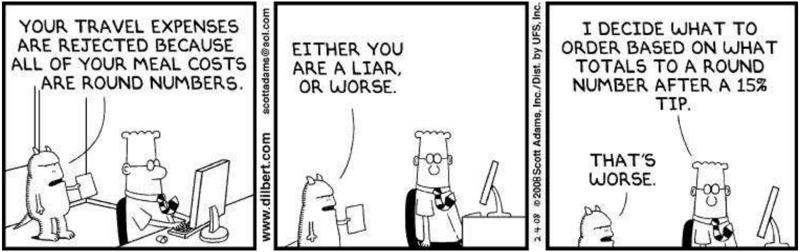
- Not all fraud can be detected by Benford's Law.

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 $x = B^y$.
  - ⋄ 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.
  - ⋄ Example: π 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.

- 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

Clicker Question

First 60 numbers of the form 2^n

digit	# Obs	# Pred	Obs Prob	Benf Prob
1	18	18.1	.300	.301
2	12	10.6	.200	.176
3	6	7.5	.100	.125
4	6	5.8	.100	.097
5	6	4.8	.100	.079
6	4	4.0	.067	.067
7	2	3.5	.033	.058
8	5	3.1	.083	.051
9	1	2.7	.017	.046

First 60 numbers of the form 2^n

digit	# Obs	# Pred	Obs Prob	Benf Prob
1	18	18.1	.300	.301
2	12	10.6	.200	.176
3	6	7.5	.100	.125
4	6	5.8	.100	.097
5	6	4.8	.100	.079
6	4	4.0	.067	.067
7	2	3.5	.033	.058
8	5	3.1	.083	.051
9	1	2.7	.017	.046

As $N \to \infty$, is $\{2^n\}_{n=0}^{N}$ Benford? (a) yes (b) no (c) open.

First 60 numbers of the form 2^n

digit	# Obs	# Pred	Obs Prob	Benf Prob
1	18	18.1	.300	.301
2	12	10.6	.200	.176
3	6	7.5	.100	.125
4	6	5.8	.100	.097
5	6	4.8	.100	.079
6	4	4.0	.067	.067
7	2	3.5	.033	.058
8	5	3.1	.083	.051
9	1	2.7	.017	.046

As $N \to \infty$, is $\{2^n\}_{n=0}^{N}$ Benford? (a) yes (b) no (c) open.

Are the 9s low in limit? (a) yes (b) no (c) open.

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

Statement

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

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

Data Analysis

- χ^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 χ^2, more likely correct model.

- Will study γ^n, e^n, π^n.

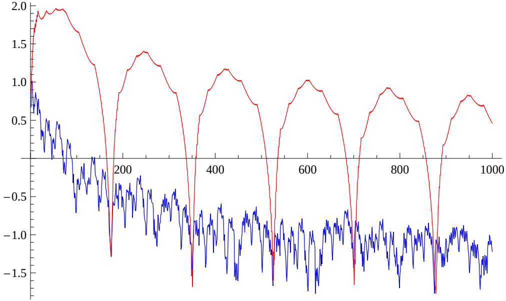
Logarithms and Benford's Law

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

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

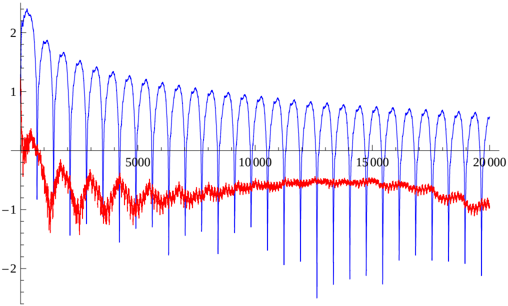
Logarithms and Benford's Law: Base 10

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



Logarithms and Benford's Law: Base 20

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



General Theory

Mantissas

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

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

Key observation: $\log_{10}(x) = \log_{10}(\widetilde{x})$ mod 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 $b - a$:

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

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 $b - a$:

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

- 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 $b-a$:

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

- 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

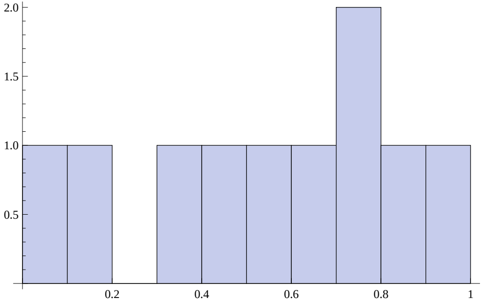
Equidistribution

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

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

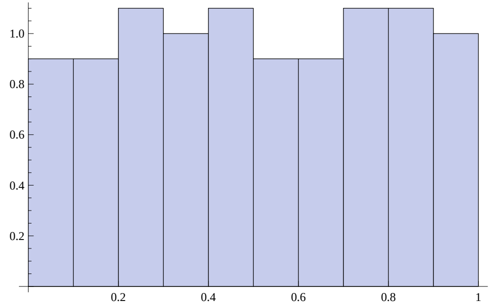
- Thm: $\beta \notin \mathbb{Q}$, $n\beta$ is equidistributed mod 1.
- Examples: $\log_{10} 2, \log_{10}\left(\frac{1+\sqrt{5}}{2}\right) \notin \mathbb{Q}$.
  Proof: if rational: $2 = 10^{p/q}$.
  Thus $2^q = 10^p$ or $2^{q-p} = 5^p$, impossible.

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



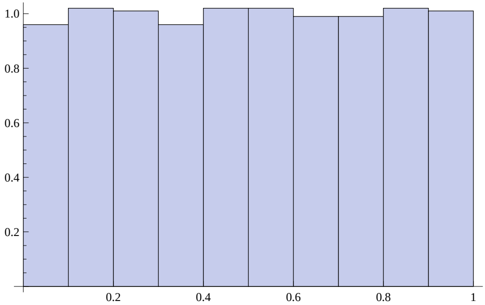
$$n\sqrt{\pi} \bmod 1 \text{ for } n \leq 10$$

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



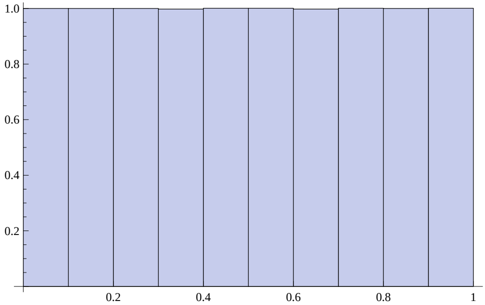
$n\sqrt{\pi} \bmod 1$ for $n \leq 100$

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



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

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

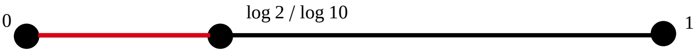
Fundamental Equivalence

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

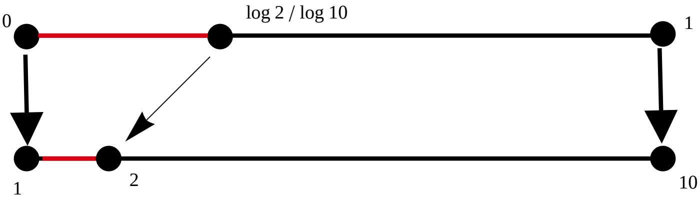
Data set $\{x_i\}$ is Benford base B if $\{y_i\}$ is equidistributed mod 1, where $y_i = \log_B x_i$.



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 $d \leq M_B(x) < d + 1$.
- $\log_B d \leq y < \log_B(d+1)$, $y = \log_B x \bmod 1$.
- 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:
  - $\diamond$ $a_{n+1} = 2a_n - a_{n-1}$

Examples

- Fibonacci numbers are Benford base 10.
  $a_{n+1} = a_n + a_{n-1}$.
  Guess $a_n = n^r$: $r^{n+1} = r^n + r^{n-1}$ 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: $a_n = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2}\right)^n$.

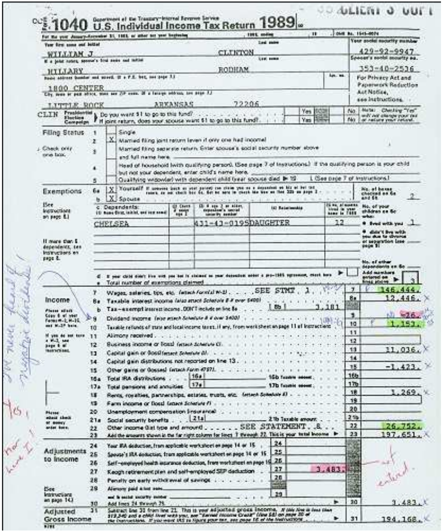
- Most linear recurrence relations Benford:
  - $\diamond$ $a_{n+1} = 2a_n - a_{n-1}$
  - $\diamond$ take $a_0 = a_1 = 1$ or $a_0 = 0$, $a_1 = 1$.

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



1040 U.S. Individual Income Tax Return 1989

CLIENT'S COPY

WILLIAM J CLINTON 429-92-9947

HILLARY RODHAM 353-40-2536

1800 CENTER

LITTLE ROCK ARKANSAS 72206

Filing Status

Exemptions

CHELSEA 431-43-0195 DAUGHTER 12

Income

7 Wages, salaries, tips, etc. SEE STMT 1 146,444.

8a Taxable interest income 12,446.

8b Tax-exempt interest income 3,181

9 Dividend income 26.

10 Taxable refunds of state and local income taxes 1,153.

13 Capital gain or (loss) 11,036.

15 Other gains or (losses) -1,423.

18 Rents, royalties, partnerships, estates, trusts, etc. 1,269.

22 Other income SEE STATEMENT 8 26,752.

23 Add the amounts shown in the far right column for lines 7 through 22. This is your total income 197,651.

Adjustments to Income

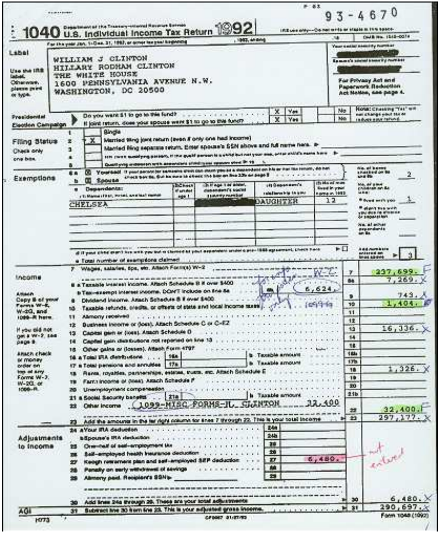
27 Keogh retirement plan and self-employed SEP deduction 3,483.

30 Add lines 24 through 29 3,483.

Adjusted Gross Income

31 Subtract line 30 from line 22. This is your adjusted gross income. 194,168.

Applications for the IRS: Detecting Fraud


93-4670

1040 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

Filing Status

Exemptions

CHELSEA [redacted] DAUGHTER 12

Income

7 Wages, salaries, tips, etc. Attach Form(s) W-2	237,699.
8a Taxable interest income	7,269.
9 Dividend income	743.
10 Taxable refunds, credits, or offsets of state and local income taxes	1,404.
13 Capital gain or (loss). Attach Schedule D	16,336.
18 Rents, royalties, partnerships, estates, trusts, etc. Attach Schedule E	1,326.
22 Other income 1099-MISC FORMS-H. CLINTON	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

27 Keogh retirement plan and self-employed SEP deduction	6,480.
30 Add lines 24a through 29. These are your total adjustments	6,480.
31 Subtract line 30 from line 23. This is your adjusted gross income.	290,697.

AGI

Form 1040 (1992)

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.

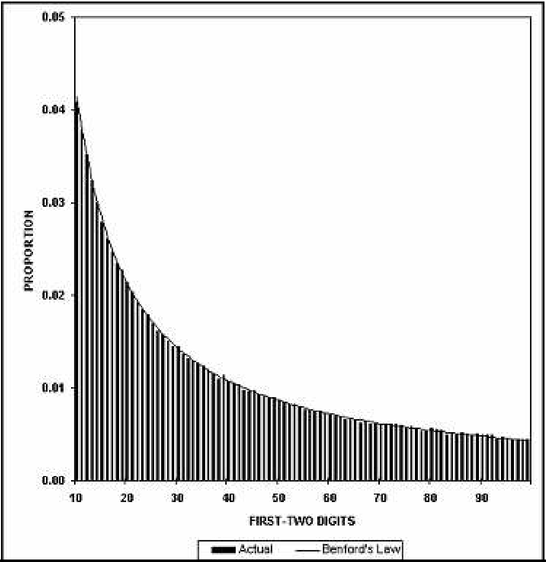
Detecting Fraud

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....

The Modulo 1 Central Limit Theorem

Needed Input: Poisson Summation Formula

Poisson Summation Formula

f nice:

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

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

Needed Input: Poisson Summation Formula

Poisson Summation Formula

f nice:

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

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

What is 'nice'?

- f Schwartz more than enough.
- f twice continuously differentiable & f, f', f'' decay like $x^{-(1+\eta)}$ for an $\eta > 0$ (g decays like x^{-a} if $\exists x_0, C$ st $|x| > x_0$, $|g(x)| \leq C/|x|^a$).

Modulo 1 Central Limit Theorem

The Modulo 1 Central Limit Theorem for Independent

Let $\{Y_m\}$ be independent continuous random variables on $[0, 1)$, not necessarily identically distributed, with densities $\{g_m\}$. A necessary and sufficient condition for $Y_1 + \cdots + Y_M$ modulo 1 to converge to the uniform distribution as $M \to \infty$ (in $L_1([0, 1])$) is that for each $n \neq 0$ we have $\lim_{M\to\infty} \widehat{g_1}(n) \cdots \widehat{g_M}(n) = 0$.

Modulo 1 Central Limit Theorem

The Modulo 1 Central Limit Theorem for Independent

Let $\{Y_m\}$ be independent continuous random variables on $[0,1)$, not necessarily identically distributed, with densities $\{g_m\}$. A necessary and sufficient condition for $Y_1 + \cdots + Y_M$ modulo 1 to converge to the uniform distribution as $M \to \infty$ (in $L_1([0,1])$) is that for each $n \neq 0$ we have $\lim_{M\to\infty} \widehat{g_1}(n) \cdots \widehat{g_M}(n) = 0$.

Application to Benford's law: If $X = X_1 \cdots X_M$ then

$$\log_{10} X \ = \ \log_{10} X_1 + \cdots + \log_{10} X_M \ := \ Y_1 + \cdots + Y_M.$$

Products of Random Variables and the Fourier Transform

Preliminaries

- $X_1 \cdots X_n \Leftrightarrow Y_1 + \cdots + Y_n$ mod 1, $Y_i = \log_B X_i$
- Density Y_i is g_i, density $Y_i + Y_j$ is

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

- $h_n = g_1 * \cdots * g_n$, $\widehat{g}(\xi) = \widehat{g}_1(\xi) \cdots \widehat{g}_n(\xi)$.
- Dirac delta functional: $\int \delta_\alpha(y) g(y) dy = g(\alpha)$.

Fourier input

- Fejér kernel:

$$F_N(x) = \sum_{n=-N}^{N} \left(1 - \frac{|n|}{N}\right) e^{2\pi i n x}.$$

- Fejér series:

$$T_N f(x) = (f * F_N)(x) = \sum_{n=-N}^{N} \left(1 - \frac{|n|}{N}\right) \widehat{f}(n) e^{2\pi i n x}.$$

- Lebesgue's Theorem: $f \in L^1([0,1])$. As $N \to \infty$, $T_N f$ converges to f in $L^1([0,1])$.
- $T_N(f * g) = (T_N f) * g$: convolution assoc.

Modulo 1 Central Limit Theorem

Theorem (M– and Nigrini 2007)

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

Generalizations

- Levy proved for i.i.d.r.v. just one year after Benford's paper.

- Generalized to other compact groups, with estimates on the rate of convergence.
  $\diamond$ Stromberg: n-fold convolution of a regular probability measure on a compact Hausdorff group G converges to normalized Haar measure in weak-star topology iff support of the distribution not contained in a coset of a proper normal closed subgroup of G.

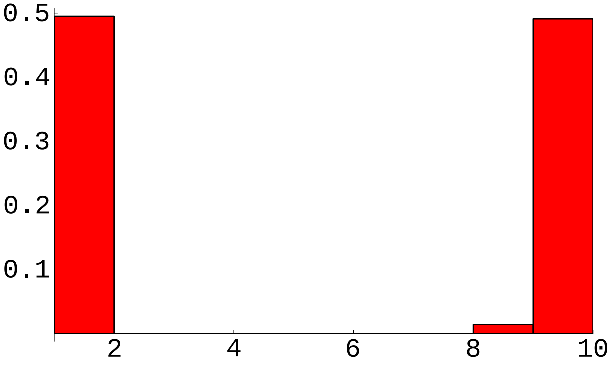
Theorem (M– and Nigrini 2007)

$\{Y_m\}$ indep. discrete random variables on $[0,1)$, not necc. identically distributed, densities

$$g_m(x) = \sum_{k=1}^{r_m} w_{k,m}\delta_{\alpha_{k,m}}(x),\, w_{k,m} > 0, \sum_{k=1}^{r_m} w_{k,m} = 1.$$

Assume that there is a finite set $A \subset [0,1)$ such that all $\alpha_{k,m} \in A$. $Y_1 + \cdots + Y_M \bmod 1$ converges weakly to the uniform distribution as $M \to \infty$ iff $\forall n \neq 0$, $\lim_{M\to\infty} \widehat{g_1}(n)\cdots\widehat{g_M}(n) = 0$.

Distribution of digits (base 10) of 1000 products $X_1 \cdots X_{1000}$, where $g_{10,m} = \phi_{11^m}$.
$\phi_m(x) = m$ if $|x - 1/8| \leq 1/2m$ (0 otherwise).



Proof of Modulo 1 CLT

- Density of sum is $h_\ell = g_1 * \cdots * g_\ell$.
- Suffices show $\forall \epsilon$: $\lim_{M\to\infty} \int_0^1 |h_M(x) - 1| dx < \epsilon$.
- Lebesgue's Theorem: N large,

$$||h_1 - T_N h_1||_1 = \int_0^1 |h_1(x) - T_N h_1(x)| dx < \frac{\epsilon}{2}.$$

- Claim: above holds for h_M for all M.

Proof of claim

$$T_N h_{M+1} = T_N(h_M * g_{M+1}) = (T_N h_M) * g_{M+1}$$

$$\begin{aligned}
||h_{M+1} - T_N h_{M+1}||_1 &= \int_0^1 |h_{M+1}(x) - T_N h_{M+1}(x)|dx \\
&= \int_0^1 |(h_M * g_{M+1})(x) - (T_N h_M) * g_{M+1}(x)|dx \\
&= \int_0^1 \left| \int_0^1 (h_M(y) - T_N h_M(y))\, g_{M+1}(x-y) \right| dydx \\
&\leq \int_0^1 \int_0^1 |h_M(y) - T_N h_M(y)| g_{M+1}(x-y) dxdy \\
&= \int_0^1 |h_M(y) - T_N h_M(y)| dy \cdot 1 \; < \; \frac{\epsilon}{2}.
\end{aligned}$$

Proof of Modulo 1 CLT (continued)

Show $\lim_{M\to\infty} ||h_M - 1||_1 = 0$.
Triangle inequality:

$$||h_M - 1||_1 \le ||h_M - T_N h_M||_1 + ||T_N h_M - 1||_1.$$

Choices of N and ϵ:

$$||h_M - T_N h_M||_1 < \epsilon/2.$$

Show $||T_N h_M - 1||_1 < \epsilon/2$.

$$||T_N h_M - 1||_1 = \int_0^1 \left| \sum_{\substack{n=-N \\ n\neq 0}}^{N} \left(1 - \frac{|n|}{N}\right) \widehat{h_M}(n) e^{2\pi i n x} \right| dx$$

$$\leq \sum_{\substack{n=-N \\ n\neq 0}}^{N} \left(1 - \frac{|n|}{N}\right) |\widehat{h_M}(n)|$$

$\widehat{h_M}(n) = \widehat{g}_1(n) \cdots \widehat{g_M}(n) \longrightarrow_{M\to\infty} 0.$
For fixed N and ϵ, choose M large so that $|\widehat{h_M}(n)| < \epsilon/4N$ whenever $n \neq 0$ and $|n| \leq N$.