# Benford's Law and Fraud Detection, or: Why the IRS Should Care About Number Theory!

Steven J Miller
Williams College

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

Bronfman Science Lunch
Williams College, October 21, 2008

## Summary

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

## Caveats!

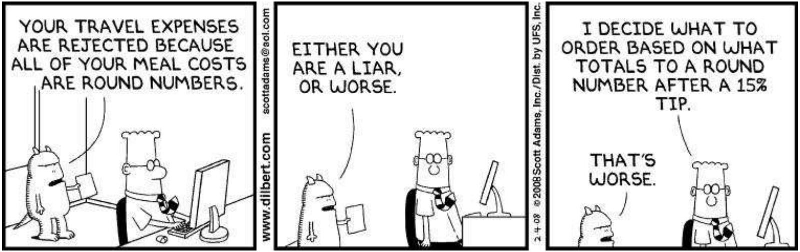
- Not all fraud can be detected by Benford's Law.

## 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.



## 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 $b-a$:

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

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

## Equidistribution and Benford's Law

### Equidistribution

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

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

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

## Equidistribution and Benford's Law

### Equidistribution

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

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

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

## Equidistribution and Benford's Law

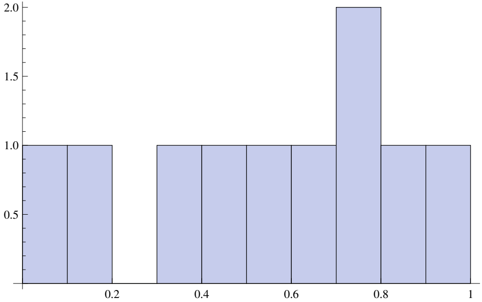
### Equidistribution

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

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

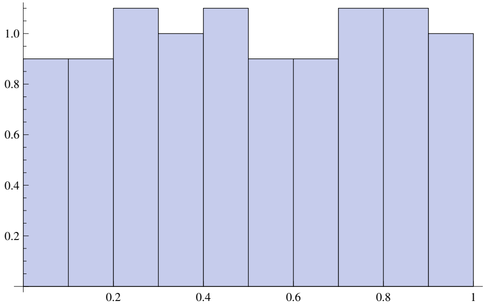
- Thm: $\beta \notin \mathbb{Q}$, $n\beta$ is equidistributed mod 1.
- Examples: $\log_{10} 2, \log_{10}\left(\frac{1+\sqrt{5}}{2}\right) \notin \mathbb{Q}$.
  *Proof:* if rational: $2 = 10^{p/q}$.
  Thus $2^q = 10^p$ or $2^{q-p} = 5^p$, impossible.

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



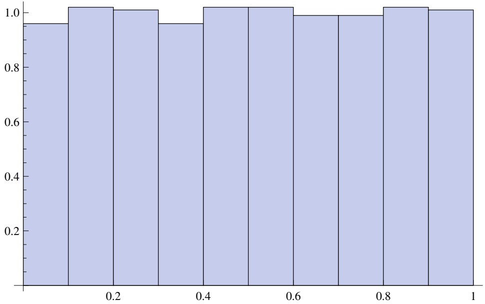

$n\sqrt{\pi} \bmod 1$ for $n \leq 10$

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



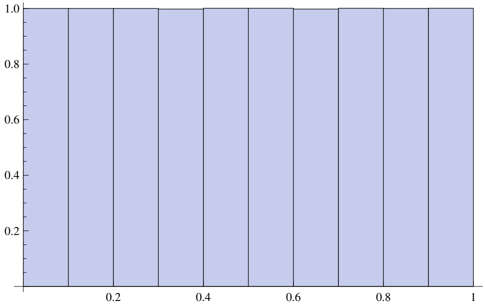

$n\sqrt{\pi} \bmod 1$ for $n \leq 100$

# Example of Equidistribution: $n\sqrt{\pi}$ 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 $n \leq 10,000$

## Logarithms and Benford's Law

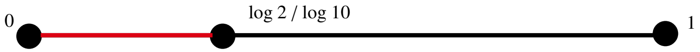
### Fundamental Equivalence

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

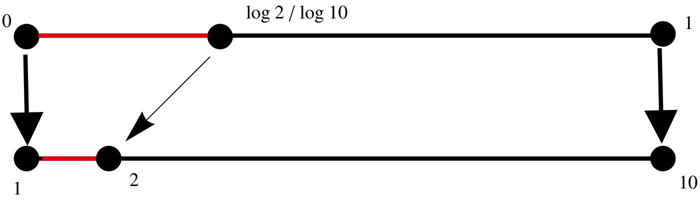
Data set $\{x_i\}$ is Benford base $B$ if $\{y_i\}$ is equidistributed mod 1, where $y_i = \log_B x_i$.



## Logarithms and Benford's Law

### Fundamental Equivalence

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



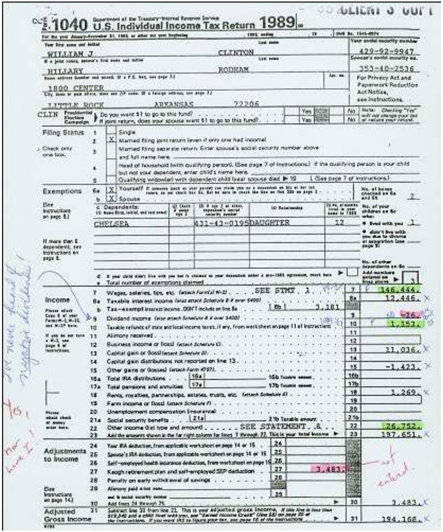
## Examples

- $2^n$ is Benford base 10 as $\log_{10} 2 \notin \mathbb{Q}$.
- 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: $a_n = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2}\right)^n$.

# Applications

# Applications for the IRS: Detecting Fraud


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

CLIENT'S COPY

Your first name and initial: WILLIAM J — Last name: CLINTON — Your social security number: 429-92-9947
If a joint return, spouse's first name and initial: HILLARY — Last name: RODHAM — Spouse's social security no.: 353-40-2536
Home address (number and street): 1800 CENTER
City, town or post office, state and ZIP code: 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? Yes / No. If joint return, does your spouse want $1 to go to this fund? Yes / No. Note: Checking "Yes" will not change your tax or reduce your refund.

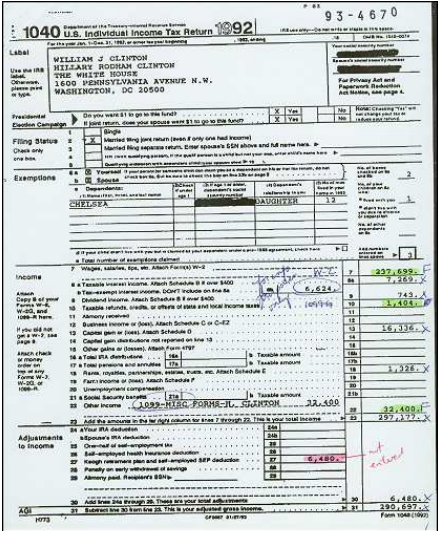
Filing Status (Check only one box.)
1 Single
2 X Married filing joint return (even if only one had income)
3 Married filing separate return. Enter spouse's social security number above and full name here.
4 Head of household (with qualifying person). (See page 7 of Instructions.) If the qualifying person 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 — No. of boxes checked on 6a and 6b: 2
c Dependents: CHELSEA — 431-43-0195 — DAUGHTER — 12 — No. of your children on 6c who: lived with you 1
e Total number of exemptions claimed — Add numbers entered on lines above ▶ 3

| Line | Income | Amount |
| --- | --- | --- |
| 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] 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 / 16b Taxable amount | |
| 17a | Total pensions and annuities / 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 / 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. |

| Line | Adjustments to Income | Amount |
| --- | --- | --- |
| 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. |
| 31 | Subtract line 30 from line 23. This is your adjusted gross income. ▶ | 194,168. |

# Applications for the IRS: Detecting Fraud


93-4670

1040 U.S. Individual Income Tax Return 1992

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

Income
7 Wages, salaries, tips, etc. Attach Form(s) W-2 — W-2 — 237,699.
8a Taxable interest income. Attach Schedule B if over $400 — 7,269.
b Tax-exempt interest income. DON'T include on line 8a — 6,624
9 Dividend income. Attach Schedule B if over $400 — 743.
10 Taxable refunds, credits, or offsets of state and local income taxes — 1,404.
11 Alimony received
12 Business income or (loss). Attach Schedule C or C-EZ — 16,336.
13 Capital gain or (loss). Attach Schedule D
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.

not entered

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

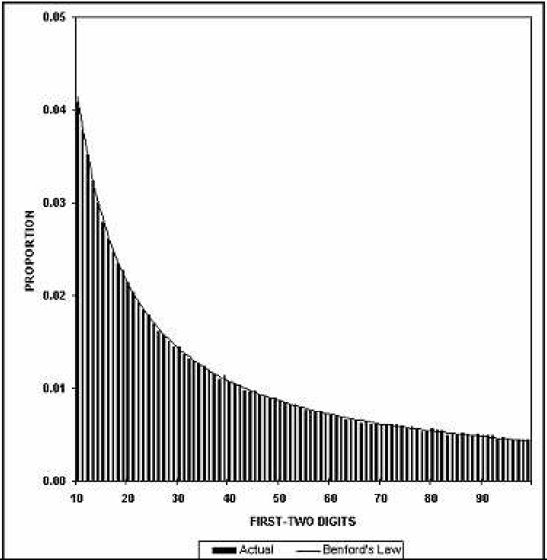
- Audit of a bank revealed huge spike of numbers starting with 48 and 49, most due to one person.
- Write-off limit of $5,000. Officer had friends applying for credit cards, ran up balances just under $5,000 then he would write the debts off.

## Detecting Fraud

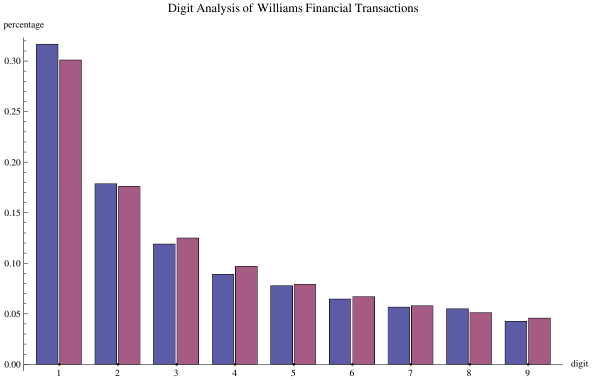
### Enron

- 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.

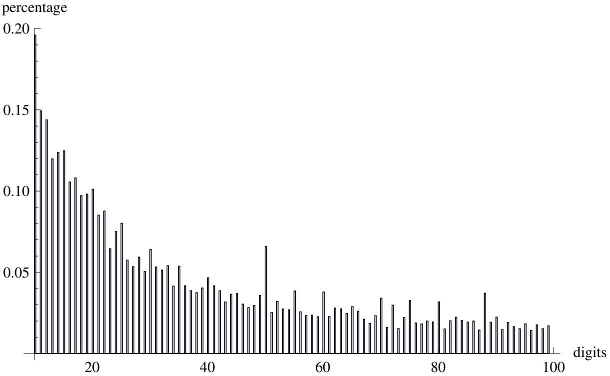
# Data Integrity: Stream Flow Statistics: 130 years, 457,440 records



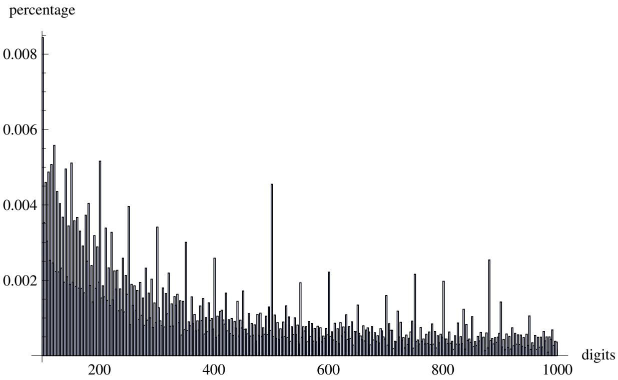
# Analysis of Williams College Transactions (thanks to Richard McDowell): September 6, 2006 to June 29, 2007: 64,000+ transactions



# Analysis of Williams College Transactions (thanks to Richard McDowell): September 6, 2006 to June 29, 2007: 64,000+ transactions



## Analysis of Williams College Transactions (thanks to Richard McDowell): September 6, 2006 to June 29, 2007: 64,000+ transactions



# Conclusions

## Conclusions and Future Investigations

- Diverse systems exhibit Benford behavior.

## Conclusions and Future Investigations

- Diverse systems exhibit Benford behavior.

- Ingredients of proofs (logarithms, equidistribution).

## Conclusions and Future Investigations

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

## Conclusions and Future Investigations

- Diverse 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.