Newcomb, Benford, Pareto, Heaps, and Zipf
Are arbitrary numbers random?

Nelson H. F. Beebe

Research Professor
University of Utah
Department of Mathematics, 110 LCB
155 S 1400 E RM 233
Salt Lake City, UT 84112-0090
USA

Email: beebe@math.utah.edu, beebe@acm.org,
beebe@computer.org (Internet)
WWW URL: http://www.math.utah.edu/~beebe
Telephone: +1 801 581 5254
FAX: +1 801 581 4148

10 December 2011 18 January 2012 20 February 2013
27 January 2016

Numbers and distributions

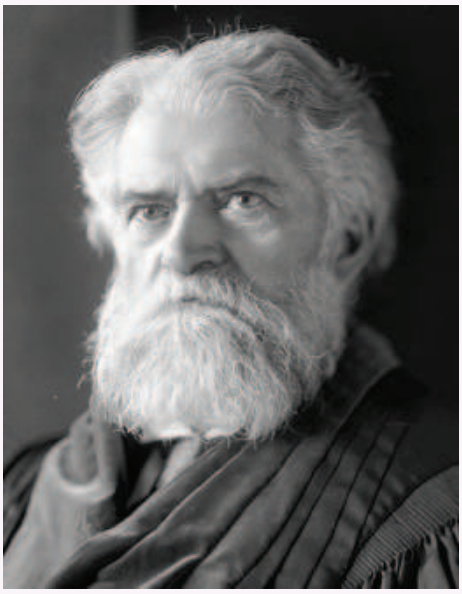
Simulations usually need a source of numer
are sometimes a suitable source.

However, random numbers may conform to
normal, exponential, logarithmic, Poisson, ..

The key question is:

Do numbers in real data match a

A negative answer



Simon Newcomb (1835–1909) Canadian / American astronomer, mathematician, economist, linguist, mountaineer

Note on the frequency of use of the different digits in natural numbers, *American Journal of Mathematics*, **4**(1–4) 39–40 (1881). The short note begins:

That the ten digits do not occur with equal frequency must be evident to any one making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones.

But wait...

[Newcomb was a co-founder of the **America**
its first President (1899–1904).]

Consider the integers from, say, 100 to 999.
100 more in $[200, 299]$, and so on up to the

We conclude that for random numbers from

leading digits have equa

There are nine such digits, $1, 2, \ldots, 9$, so th
$1/9 \approx 0.111$.

Newcomb's prediction

The law of probability of the occurrence of numbers is such that all mantissæ of their logarithms are equally probable.

digit	first	second
0		0.1197
1	0.3010	0.1139
2	0.1761	0.1088
3	0.1249	0.1043
4	0.0969	0.1003
5	0.0792	0.0967
6	0.0669	0.0934
7	0.0580	0.0904
8	0.0512	0.0876
9	0.0458	0.0850

In the case of the third figure the probability will be nearly the same for each digit, and for the fourth and following ones the difference will be inappreciable.

Newcomb's conclusion

It is curious to remark that this law w whether a large collection of indepen were composed of natural numbers o

Then Newcomb's work was forgotten for 57

Benford's rediscovery

American physicist **Frank Benford** (1883–1948), in **The Law of Anomalous Numbers**, *Proceedings of the American Philosophical Society*, **78**(4) 551–572, March (1938), perhaps unaware of Newcomb's work (but he mentions the *dirty pages* phenomenon), rediscovered the same curiosity.

Benford's paper was noticed, and the law is named after him.

[photograph ca. 1912, age 29]

Benford's rediscovery [continue

Benford illustrated the phenomenon with a g

river (drainage?) areas	$1/n$, $\sqrt{n}$
land area	design d
US population	*Reader's*
physical constants	cost data
newspaper items	X-ray vol
specific heats	America
pressure lost in air flow	black-bo
H.P. lost in air flow	AMS stre
drainage	n^1, n^2, n
atomic & molecular weights	death ra
house numbers	river drai

He gave frequency data for each, and a cum frequencies: **0.306, 0.185, 0.124, 0.094, 0.0 0.047**.

Why Benford got a Law, and Newcomb did not

Benford gave much more data, and provided more mathematical arguments, in support of his **Law of Anomalous Numbers**, than Newcomb did in 1881.

> *Benford's paper was published in 1938 in a journal of rather limited circulation and not usually read by mathematicians. It so happened that it was immediately followed in the same issue by a physics paper which became of some importance for secret nuclear work during World War II. That is why Benford's paper caught the attention of physicists in the early 1940's and was much discussed.*
>
> ***Jonothan L. Logan and Samuel A. Goudsmit**, The First Digit Phenomenon, Proceedings of the American Philosophical Society, **122**(4) 193–197, 18 August (1978).*

Boring and Raimi uncover New

Newcomb is briefly cited by **Edwin G. Borin**
Law of Error in Mental Measurement, *The Psychology*, **31**(1) 1–33 (1920), but only abo transcendental numbers.
Newcomb's work seems to have been uncov
The first digit problem, *American Mathema* August 1976, 95 years later. Raimi wrote:

> *This assertion, whatever it may mean, u* ***Law*** *because it has been thought by ma originated with the General Electric Co Benford. . . . There is ample precedent theorems for persons other than their di analysis would be named after Euler. B implied that the observation giving rise old one in his day. One would hate to cl now only to find later that another chan*

Benford's Law for first digits

The frequency of the first digit [in measured data] follows closely the logarithmic relation:

$$F_a = \log_{10}\left(\frac{a+1}{a}\right), \qquad \textit{Benford's original,}$$
$$= \log_{10}(1 + 1/a), \qquad \textit{modern form.}$$

Here, a is a *nonzero* leading decimal digit $1, 2, \ldots, 9$.
Benford's leading-digit frequencies are *identical* to those in Newcomb's table: **0.301, 0.176, 0.125, 0.097, 0.079, 0.067, 0.058, 0.051**, and **0.046**.
The partial sums produce cumulative frequencies given by

$$C_a = \log_{10}(1 + a)$$

with these approximate values: **0.301, 0.477, 0.602, 0.699, 0.778, 0.845, 0.903, 0.954**, and **1**. Thus, 60% start with 1, 2, or 3.

Benford's Law for second digits

For a number beginning with decimal digits

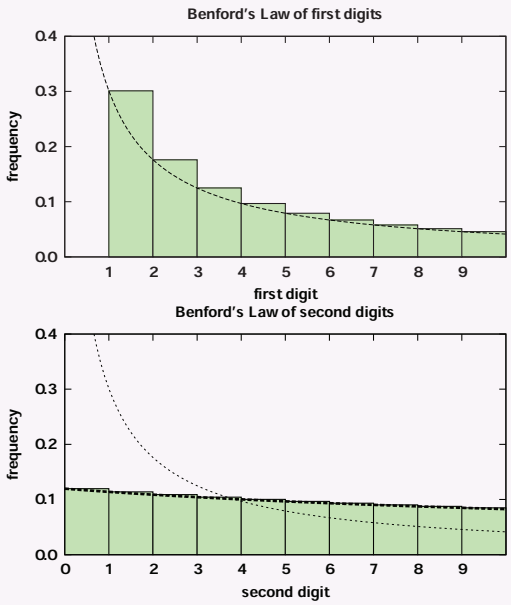
$$F_b = \frac{\log_{10}(\frac{ab+}{ab}}{\log_{10}(\frac{a+1}{a}}$$

Here, b may be any of $0, 1, 2, \ldots, 9$.

Summed over all possible leading digits, the **0.120, 0.114, 0.109, 0.104, 0.100, 0.097, 0.**

In some cases, second-digit data have prove data, and have been used for examining ele fraud (e.g., Argentina, Germany, Iran, Puerto

Benford's Law pictorially



Benford's Law for arbitrary digi

For a number beginning with decimal digits a

$$\log(1+x) \approx (x - x^2/2 + x^3/3 - x^4/4 -$$
$$\log_{10}(1+x) \approx (x - x^2/2 + x^3/3 - x^4/4 -$$
$$F_q = \frac{\log_{10}(\frac{abc\cdots opq+1}{abc\cdots opq})}{\log_{10}(\frac{abc\cdots op+1}{abc\cdots op})} = \frac{\log_{10}(}{\log_{10}(}$$
$$\to 1/10,$$

For example, if $abcdefgh = 12345678$, then

$$F_9 = \frac{\log_{10}(\frac{123456789+1}{123456789})}{\log_{10}(\frac{12345678+1}{12345678})} \approx 0.0$$

Thus, after the first few leading digits, **there frequencies**.

Computational note: **use** `log1p(x)` **instead**

Benford's Law and percentage growth

Consider a company with $1,000,000 revenues:

- Leading digit of 1: income increases by 100% to $2,000,000.
- Leading digit of 2: income increases by 50% to $3,000,000.
- Leading digit of 3: income increases by 33% to $4,000,000.
- Leading digit of 4: income increases by 25% to $5,000,000.
- ...
- Leading digit of 9: income increases by 11% to $10,000,000.

Suggestion: **If percentage growth is roughly constant, then smaller leading digits should be more common.**

Growth is more likely to be *geometric* than *arithmetic*.

Frequencies decrease [0.353, 0.177, 0.118, 0.088, 0.071, 0.059, 0.050, 0.044, and 0.039] but do not match Benford's Law.

Benford's Law: two observation

Benford's 'law of first digits' has a histor decades and has produced a literature that it shows a lack of understanding th **fundamental and general** *rather than of a particular data set.*

***B. K. Jones**, Logarithmic distributions in reli Microelectronics Reliability **42**(4–5) 779–786 (2*

Wallace (2002) suggests that if the mea numbers is larger than the median and positive, the data set likely follows a Be follows that **the larger the ratio of the median, the more closely the set will**

***C. Durtschi et al.**, The Effective Use of Benfo Detecting Fraud in Accounting Data, Journal 17–34 (2004).*

Benford's Law and mixed distributions

If distributions are selected at random (in any "unbiased way") and random samples are take from these distributions, then the significant-digit frequencies of the **combined sample** *will converge to Benford's distribution, even though the individual distributions selected may not closely follow the law.*

***Theodore P. Hill**, The First Digit Phenomenon, American Scientist, **86**(4) 358–363 July / August (1998).*

Benford's Law in other number

If Benford's Law holds for decimal numbers, number bases, provided that those bases ar the base b in the logarithms in the digit-frequ For example,

Digit	0	1	2	3	
F_a		1.000			
F_b	0.585	0.415			
F_a		0.500	0.292	0.208	
F_b	0.304	0.261	0.230	0.206	
F_a		0.333	0.195	0.138	0.
F_b	0.151	0.141	0.133	0.126	0.

See **Theodore Hill**, **Base-invariance impli** *Proceedings of the American Mathematical* March 1995.

Benford's Law observed in real data

Digit	0	1	2	3	4	5	6	7	8	9
							1990 US Census data (5148 values)			
F_a		0.298	0.215	0.113	0.082	0.098	0.056	0.055	0.034	0.049
F_b	0.166	0.090	0.096	0.081	0.100	0.122	0.076	0.073	0.066	0.130
							Atomic weights (110 values)			
F_a		0.391	0.309	0.045	0.036	0.055	0.036	0.036	0.036	0.055
F_b	0.173	0.045	0.109	0.100	0.145	0.145	0.055	0.055	0.091	0.082
							Country areas (1505 values)			
F_a		0.312	0.275	0.100	0.067	0.058	0.062	0.046	0.029	0.050
F_b	0.167	0.221	0.092	0.092	0.062	0.067	0.075	0.067	0.083	0.075
							Country population (163 values)			
F_a		0.301	0.202	0.092	0.135	0.055	0.067	0.055	0.037	0.055
F_b	0.147	0.153	0.110	0.098	0.098	0.123	0.086	0.043	0.049	0.092
							Infant mortality (208 values)			
F_a		0.361	0.293	0.043	0.072	0.062	0.087	0.038	0.014	0.029
F_b	0.303	0.139	0.077	0.058	0.077	0.077	0.106	0.067	0.043	0.053
							IBM 2010 annual financial report (6126 values)			
F_a		0.333	0.160	0.163	0.086	0.068	0.053	0.047	0.045	0.046
F_b	0.172	0.169	0.096	0.084	0.085	0.095	0.079	0.074	0.079	0.068
Fibonacci numbers: $f(n) = f(n-1) + f(n-2); f(2) = f(1) = 1$ (9994 values)										
F_a		0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046
F_b	0.120	0.114	0.109	0.105	0.100	0.097	0.094	0.090	0.088	0.085

When does Benford's Law apply

Despite 135+ years since Newcomb's disco conditions for, and derivation of, Benford's L **Arno Berger and Theodore P. Hill**, **Benfor simple explanation in sight for mathemat** *Intelligencer*, **33**(1) 85–91 (2011).

There is general agreement that the law app distribution is **scale invariant**: if changing u number distribution unchanged, then Benfor [**Roger S. Pinkham**, **On the Distribution o** *Annals of Mathematical Statistics*, **32**(4) 122

Thus, we can

- do accounting in dollars, euros, pesos,
- measure distances in metric or nonmet
- measure areas in square furlongs, or s
- count people, couples, families, arms, f

When does Benford's Law apply? [continued]

The numbers in many mathematical sequences and physical distributions obey Benford's Law **exactly**, or at least closely, including:

- geometric sequences, and asymptotically-geometric sequences, like the *Fibonacci numbers* (1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, ...), and also the *Lucas numbers* (2, 1, 3, 4, 7, 11, 18, 29, 47, 76, ...) which obey $L(n) = L(n-1) + L(n-2)$, with $L(0) = 2$ and $L(1) = 1$;
- iterations like $x \leftarrow 3x + 1$, starting with $x = \textit{random number}$,
- powers of integers;
- logarithms of uniformly-distributed random numbers;
- prime numbers;
- reciprocals of all of the above;
- reciprocals of *Riemann zeta function* zeros;
- finite-state *Markov chains*;
- *Boltzmann–Gibbs* and *Fermi–Dirac* distributions (approximate), and *Bose–Einstein* distributions (exact).

When is Benford's Law inapplic

Sequences for which Benford's Law **does n**

- arithmetic sequences.
- random numbers from most common d
- digit subsets of irrational and transcend
- US telephone numbers (limited prefixes four digits all used);
- bounded sequences with restricted lead of week, month, or year; house number and weights); ...)

Where do Benford's Law publications appear?

About **1260** publications are listed in

http://www.math.utah.edu/pub/tex/bib/benfords-law.html

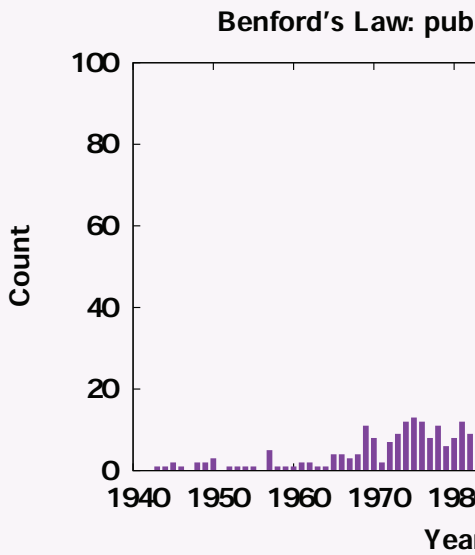
and about **980** are at

http://www.benfordonline.net/

Benford's Law articles appear in at least **465** journals in these fields:

accounting
aerobiology
auditing
astronomy
bible studies
biology
botany
business
chaos theory
chemistry
computer science
conflict resolution
criminology
demographics
drug design
earthquake detection
economics
electoral studies
engineering
finance
forensics
gambling
geography
geophysics
human resources
imaging science
library science
marketing
mathematics
medicine
networking
neuroscience
nuclear engineering
nuclear science
operations research
physics
probability
psychology
signal processing
simulation
statistics
stock-market trading
volcanology

Benford's Law literature growth



Benford's Law in accounting

Fraud and deception are common when money or politics are involved. However, many who practice in that area are unaware of Benford's Law. Their cooked data may differ sufficiently from the distribution predicted by Benford's Law that their crimes can be detected.

Several tax authorities now use Benford's Law tests in their auditing software to find tax cheats.

Fraud in numerical research data is sometimes suspected, and Benford's Law may help detect it: see **John P. A. Ioannidis**, **Why Most Published Research Findings Are False**, *PLoS Medicine*, **2**(8) 696–701, August (2005).

However, be sure first that Benford's Law is applicable, and that your statistics are good: see **Andreas Diekmann and Ben Jann**, **Benford's Law and Fraud Detection: Facts and Legends**, *German Economic Review*, **11**(3) 397–401, August (2010).

A simple test of fraud

A 200-coin-flips experiment should produce
with high probability, but few humans would

```
hoc> for (k = 1; k <= 200; ++k) pri
```

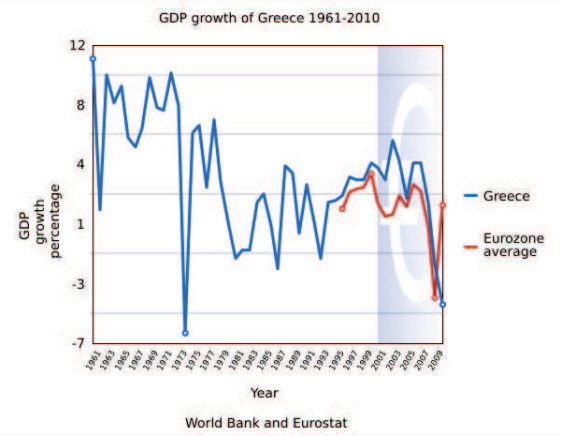
Three experiments produce (with zeros cha

```
...1.111.111..1.1.1.1111...1.1.
.1.....1..11......111.111.1....
11.1..1.11..1....1...1111111.11
1.1..11111111...111.1....11.111

1.1.1.1.11...1....1.11111......
11..1.1..1.1..1.11.11.1.11.....
..11111111..1.11....1111....1..
11111.11111111....1..11111.1.11

.11....1111.11.111.11.11.1...1.
111....11.1..11..1..1.11...1...
.1.1111...1.111.111.111.11..1..
..11.1..1..111..11.1..1.11..11.1
```


Benford's Law and the 2011 Greek debt crisis



See **Bernhard Rauch et al.**, **Fact and Fiction in EU-Governmental Economic Data**, *German Economic Review* **12**(3) 243–255, August 2011, and **Hans Christian Müller**, **How an arcane statistical law could have prevented the Greek disaster** :

http://economicsintelligence.com/2011/07/28/

How to generate data in Benfor

If a simulation involves dimensioned data wh
scale invariant, then generate starting valu

$$10^{(\text{random number u}}$$

Other distributions

Benford's Law has received wide interest and applications, but not all data conform to it.
We look briefly at four other important distributions that model real-world data.

Stigler's Law

In unpublished notes of 1945, and first prese
University of Chicago, **George J. Stigler** (*19*
Economic Sciences in Memory of Alfred Nob
distribution of leading digits arising from a m

$$F_d = \frac{1}{9}(d \ln(d) - (d+1)\ln(d+1$$

	1	2	3	4	5
Benford	0.3010	0.1761	0.1249	0.0969	0.079
Stigler	0.2413	0.1832	0.1455	0.1174	0.095

See **Joanne Lee, Wendy K. Tam Cho, and**
approach to recovering the distribution o
natural data sets, *Statistics & Probability Le*
January (2010).

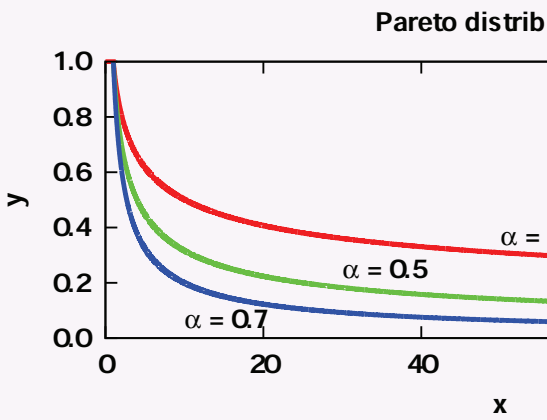
Pareto distribution

Italian economist and mathematician **Vilfredo Federico Pareto** (1848–1923) introduced the 80–20 rule in economics (80% of the wealth is owned by 20% of the people, which was true at the time in Italy, and found to be similar in other countries).
He developed the **Pareto distribution**, in which a random variable X has the property that the probability that it is greater than some number x is given by

$$\Pr(X > x) = \begin{cases} (x_m/x)^\alpha, & \text{for } x_m \le x, \\ 1, & \text{otherwise.} \end{cases}$$

The positive value x_m is a cutoff, and as $\alpha \to \infty$, the Pareto distribution approaches a *Dirac delta function*, $\delta(x - x_m)$. When this models the distribution of wealth, the exponent α is called the **Pareto index**.
Teaser: See online biographies for the relation of Pareto's economic models to the rise of Fascism in Italy in the 1920s.

Pareto distributions pictorially



Zipf's law

In 1932, American linguist **George Kingsley Zipf** (1902–1950) developed a rule that has become known as **Zipf's Law**:

> *If S is some stochastic (random) variable, the probability that S exceeds s is proportional to* $1/s$.

The variable S might be, for example, the population of a city (small cities are more numerous than large ones). See the December 2011 **National Geographic** for a story on the dramatic growth of large cities around the world.

Zipf's Law is a special case of the **Pareto distribution**.

See

http://www.nslij-genetics.org/wli/zipf/

for an online bibliography.

Heaps' law

In a 1978 book, **Information retrieval, com aspects**, **Harold Stanley Heaps** made an e linguistics that the proportion of words from exponentially with the number of words in th

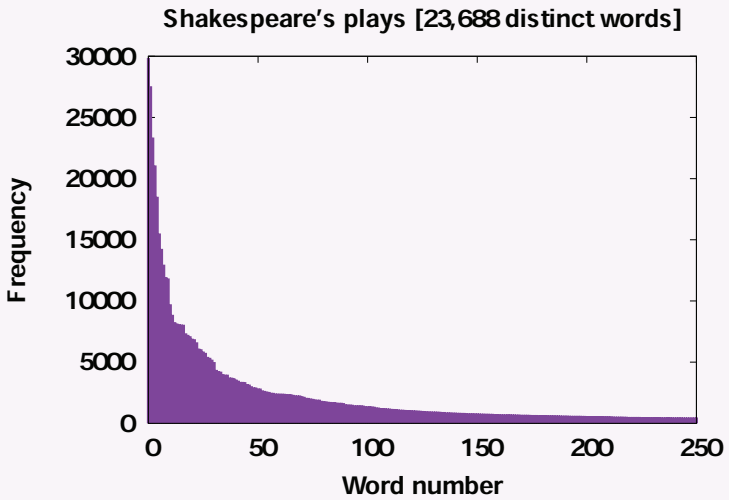
$$V_R(n) = Kn^{\beta}$$

Here, n is the text size, and K and β are em human languages, $\beta \approx 0.4$ to 0.6.
Conclusion: if $\beta < 1$, then increasing n (taki of text) results in *diminishing returns*. It is ha samples that include all, or even most, of the

Consider what **Heaps' Law** means for Web learning foreign languages, ...

Heaps' Law in Shakespeare

A small vocabulary captures most of the content, but large numbers of words are lost: 16 words (25%), 85 (50%), 250 (64%), 636 (75%), 991 (80%), 5466 (95%), 10455 (98%), 14565 (99%), ...



How to learn more

Many of the important papers on the distribu be found in

http://www.math.utah.edu/pub/tex/

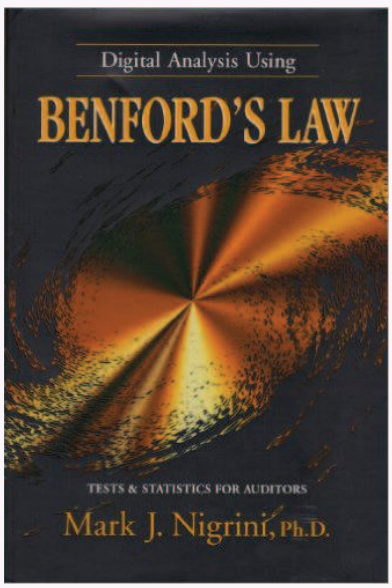
Of particular note is the survey by **Mark E. J Pareto distributions and Zipf's law**, *Conte* 323–351, September (2005),

http://dx.doi.org/10.1080/001

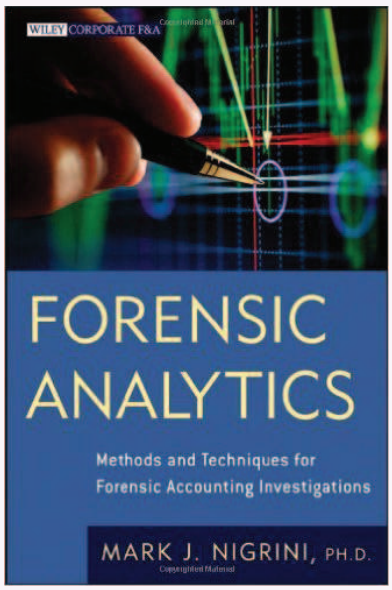
Mathematical details about the current state given by **Arno Berger and Theodore P. Hil Benford's Law**, *Probabability Surveys* **8** 1–

http://dx.doi.org/10.121

How to learn more [continued]

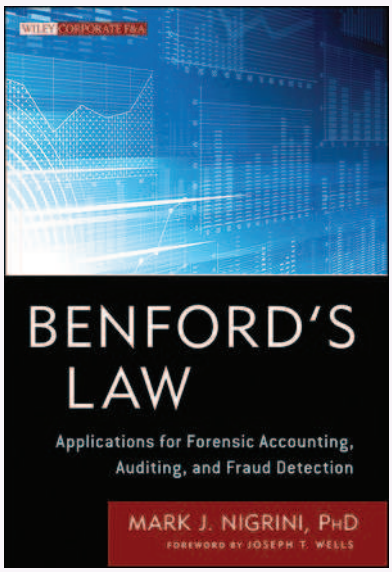


2000



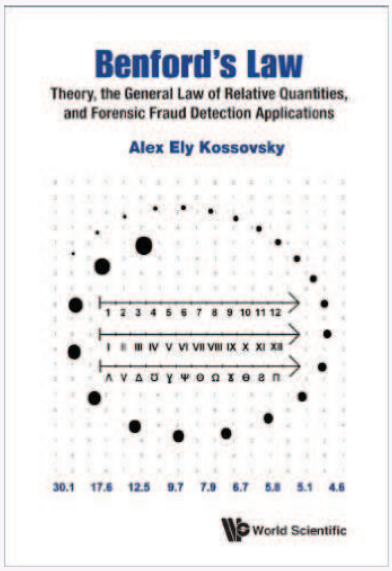
2011

How to learn more [continued]

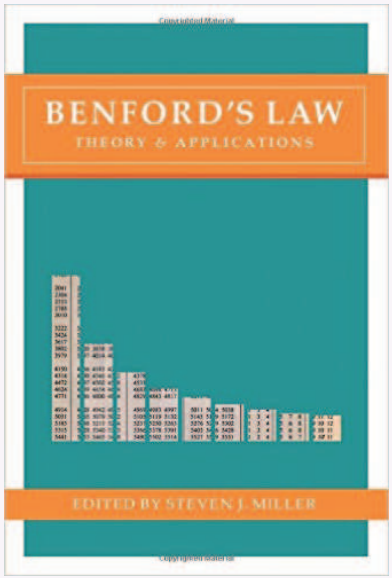


2012

How to learn more [continued]

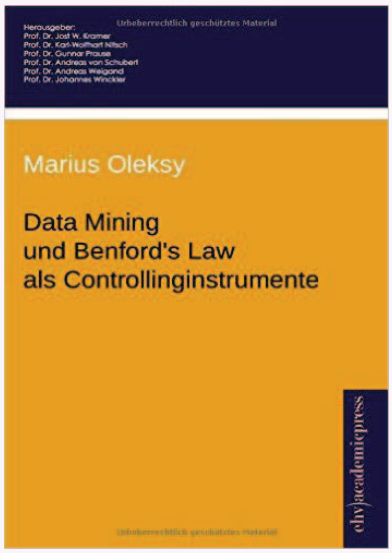


2015



2015

How to learn more [continued]



2014