Lecture 17 – Benford's Law-- Ted Hill's Theorem – and More

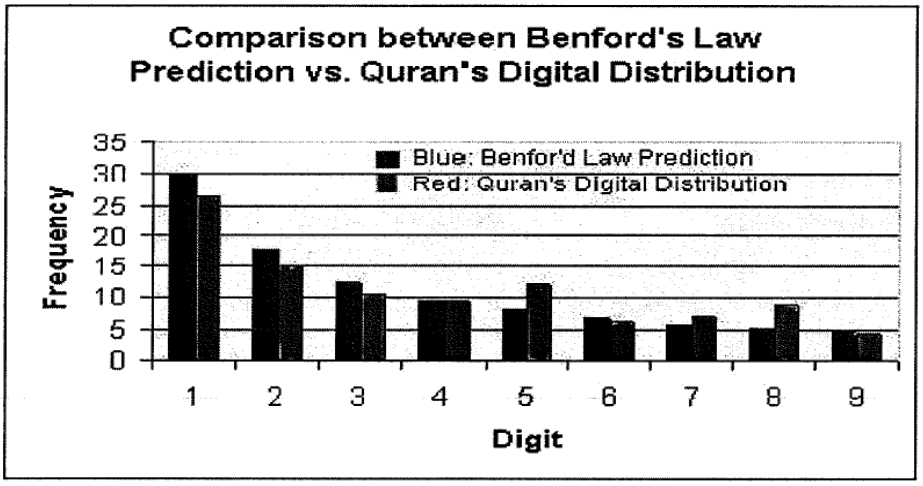
There are 9 digits (and 0) in our base 10 arithmetic. Randomly select a digit, exclusive of zero, and what are the odds the digit is 2, 7, 4, …? Since there are nine digits, you might guess 1/9 or about 11%

Here is the distribution of leading digits from different statistical series (http://testingbenfordslaw.com/):

Digit	Twitter followers	Distance of stars	UK govt spending	Spanish cities pop	Google books one-grams
1	32.62%	30.00%	29.10%	31.07%	28.32%
2	16.66%	14.67%	17.50%	18.02%	16.45%
3	11.80%	12.00%	12.20%	12.42%	13.24%
4	9.26%	10.33%	9.60%	9.18%	10.66%
5	7.63%	9.33%	8.60%	7.95%	7.88%
6	6.55%	5.67%	7.30%	6.57%	7.20%
7	5.76%	7.00%	6.10%	5/36%	5.98%
8	5.14%	7.00%	5.60%	4.95%	5.11%
9	4.56%	4.00%	4.60%	4.47%	5.16%
# records	38,670,514	300	190,379	8114	2055
min	1	4	1	5	303
max	4,706,631	3000	999,994	3,255,944	13,598,879,452
Magnitude	6	3	5	6	10

1.
2.
3.
4.

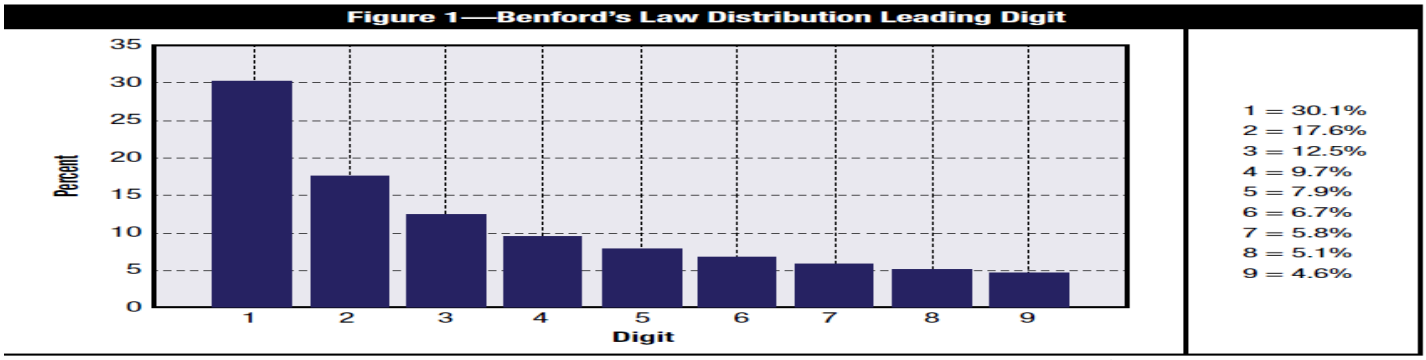
Open the Quran. Count the number of verses in the 114 sutras. You might expect again that the numbers beginning with different digits would have equal probability. This is what you get (http://www.submission.org/miracle/benford.htm):



1.What is Benford's Law of leading digits?

That the digits in any set of numbers follows a logarithmic pattern, with the first digit D having the frequency of $\log_{10}(1+1/D)$ not 1/9. There are logarithmic probabilities for other digits as well!

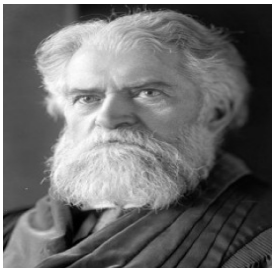
Logarithmic Scale								
1	2	3	4	5	6	7	8	9
30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%

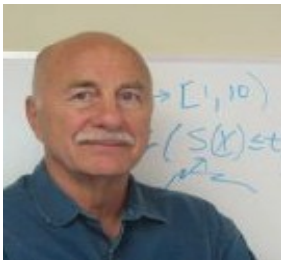


The first person to notice that the fraction of numbers starting with digit D was not 1/9th but followed a logarithmic distribution was Simon Newcomb, who noted in 1881 that log table pages were grubbier around the number 1 than the numbers 8 or 9 and fit the formula. *Newcomb, S. Note on the frequency of use of the different digits in natural numbers. American Journal of Mathematics 4(1), 39-40. ISSN:0002-9327.*

But it is called Benford's Law after Frank Benford, a physicist at GE, who rediscovered the pattern and wrote a paper in 1938 that showed that distribution of first digits of 20,229 sets of numbers from the areas of rivers to physical constants and death rates followed the law. *Benford, F. The law of anomalous numbers. Proceedings of the American Philosophical Society 78, 551-572.* In 1961 Roger Pinkham, Rutgers mathematician, claimed that any general law of digits must be scale invariant (independent of units) and that if there is such a law it must be Benford's law, but Pinkham implicitly assumed that there exists a scale-invariant probability distribution on the positive real numbers, which is not so. In 1996 Feller's classic text An Introduction to Probability claimed that "regularity and large spread implies Benford's Law" but this turns out to be false.

In the 1995 Ted Hill, of West Point, later Ga Tech, proved that data resulting from **a mix of factors** will tend to obey Benford's Law, and is "absorbing" in the sense that it causes products and other distributions that incorporate it to obey the law as well. Benford is base and scale invariance. Hill, T.P (1995) "A statistical derivation of the significant digit law" Statistical Science , 10(4), 354-363. Hill and Berger write that there is "No **Simple** Explanation In Sight For [the] Mathematical Gem", essentially because it arises from very different processes, sequences, product of Random variables, mixtures of data sets, but there are ways to understand.





Hill was a Vietnam veteran,with an amazing career in Army, academics, and he has a memoir entitled **From Beast to Berkeley** A Memoir by T.P. Hill about his wild life as a mathematician. The Benford Law is:

$$\text{Prob}\,(D_1 = d_1) = \log_{10}\left(1 + d_1^{-1}\right) \quad \text{for all } d_1 = 1, 2, \ldots, 9\,;$$

This is a probability distribution because PROBABILITY OF "D" is $\log_{10}(1 + 1/D)$ and the sum of the probabilities for the nine values = 1 bcs $\log_{10}(1 + 1/1) + \log_{10}(1 + 1/2) + \log_{10}(1 + 1/3) + \ldots + \log_{10}(1 + 1/9) = \log_{10}(2\ 3/2\ 4/3 \ldots 9/8\ 10/9) = \log_{10} 10$

But this is not all. There is a formula for the 2nd digit as well!

$$P(D_2 = d) = \sum \log_{10}(1 + (10k + d)^{-1}) \quad k = 1 \text{ to } 9.$$

The law for a set of digits is:

$$\text{Prob}((D_1, D_2, \ldots, D_n) = (d_1, d_2, \ldots, d_n)) = \log_{10}\left(1 + \left(\sum_{j=1}^{n} 10^{n-j} d_j\right)^{-1}\right)$$

The probability that the second digit is 1 is

$$\text{Prob}\,(D_2 = 1) = \sum_{j=1}^{9} \log_{10}\left(1 + \frac{1}{10j+1}\right) = \log_{10}\frac{6029312}{4638501} = 0.1138\ldots,$$

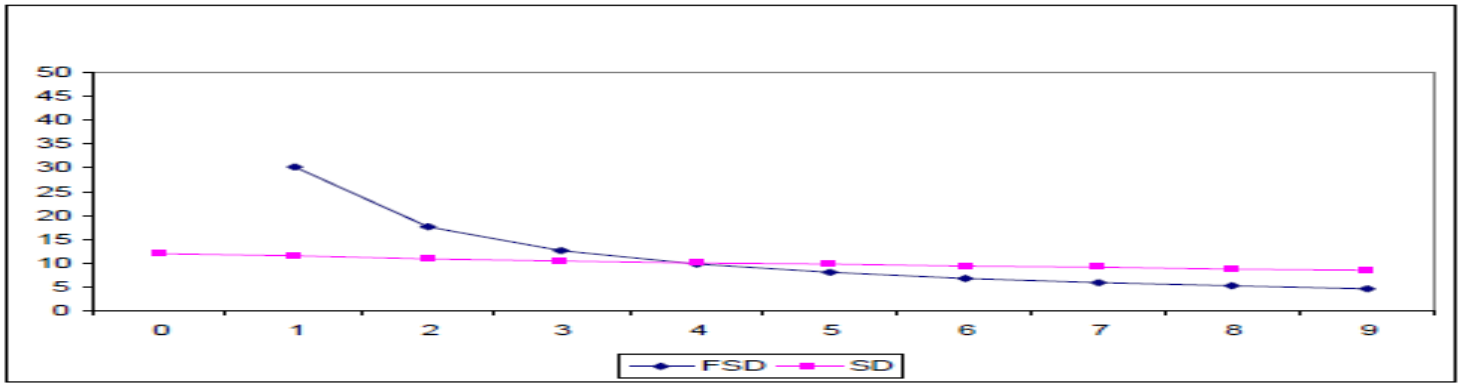
whereas, given that the first digit equals 1, the (conditional) probability that the second digit equals 1 as well is

$$\text{Prob}\,(D_2 = 1|D_1 = 1) = \frac{\log_{10} 12 - \log_{10} 11}{\log_{10} 2} = 0.1255\ldots.$$

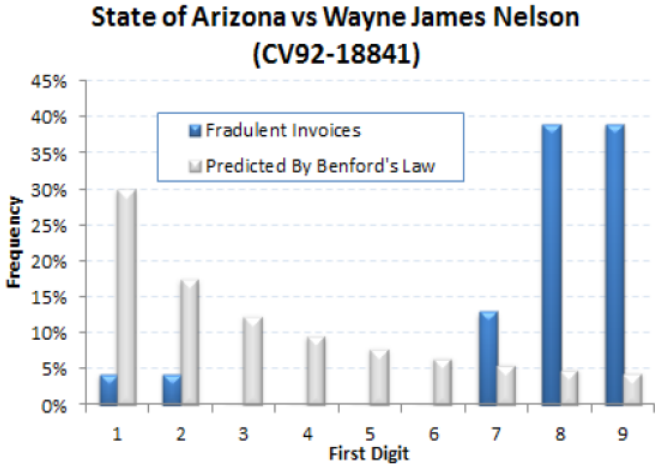
This means that the digits are not independent! The prob(311) = log10 (1+1/311) = .00139 while the Prob (319) = log10 (1+1/319) = .00136, so chance of getting a 1 in third digit after 31 is higher then chance of getting 9.

But as digits go up, the probabilities → 1/9 so the first digit deviation from 1/9 is much greater than other digits:

Figure I. Benford's Law First and Second Digits Frequencies



Accounting/business deviations often reflect judicious rounding.



Why Tether's Collapse Would Be Bad for Cryptocurrencies | WIRED Jan 30, 2017

Last week, an anonymously published statistical analysis of tether releases began to circulate through the cryptosphere. ...The report ... concluded that they violated Benford's Law—a statistical principle that in numerical data sets, more numbers tend to start with 1 than any other number, with a diminishing percentage of entries beginning with 2, 3, and so on down to 9. Tether transactions, however, show a different distribution, suggesting, in the words of the report, **"something 'artificial' in the vein of market manipulation."** The unnamed author is described in an accompanying slide presentation as a "former Googler, machine learning/statistics," who was funded by 1000x Group, a new "private community dedicated to finding the highest quality information in the crypto markets."

2.Why does it work? The magic of log growth and stopping rules --scale and base invariance

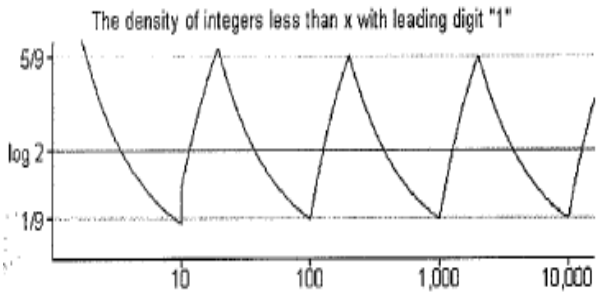
Can we get an intuitive feel for why this occurs? Logarithmic growth gives a good sense of what goes on

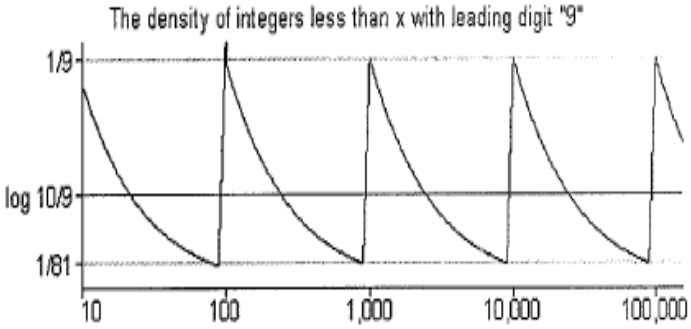
Take an index that starts at 1000 and has 20% growth per year

1000	2074	3583	1 occurrence
1200	2449	4300	1 occurrence
1440	2986	5160	1 occurrence
1738			
4 occurrences of 1	3 occurrences of 2		

It keeps growing at 20% so you get 6192, 7430, 8916 and then NO DIGIT WITH 9, 10670

Think of addresses on a street, which go from 1, 2, ...9. But on a street with 5 houses, you have addresses 1,2,3, 4, If you have a random stopping rule, you always start with 1 and get 1 but you may not get other digits. If you take the density of integers less than x you get a saw-toothed pattern.





To see the saw-tooth distribution

count numbers with digit 1 as LD along the real line

digits ≤ less than	digit	
1	1	1 of 1
2	1,2	1/2
3	1,2,3	1/3
….		
9	**1,...9**	**1/9**
10		2/10
11		3/11
….		
19		11/19 ≈ 5/9 ??
20		11/20
…		
99		**11/99= 1/9**
100		12/100
199		111/199 ≈ 5/9
999		111/999 = 1/9

so 1 has range from [1/9, 5/9)

count numbers with LD 9 along the real line

digits ≤ less than	digit
1	0/1
2	0/2
9	**1/9**
10	1/10
89	1/89 ≈ 1/81 ?? (.0112->.0123)
90	2/90
99	**11/99**
899	11/899 ≈ 1/81 (.0122)
900	12/900
999	111/999 = 1/9
8999	111/8999≈ 1/81 (.0123)

so has range (1/81, 1/9]

The math for building on these observations still puzzled people, in part because it was not clear how to define the domain for assessing the probability of getting a given first digit. The chance of getting any digit on the real line is 0, so how do you define the density?

One solution is to use the natural density of a set A: the limit as n goes to infinity of the sequence

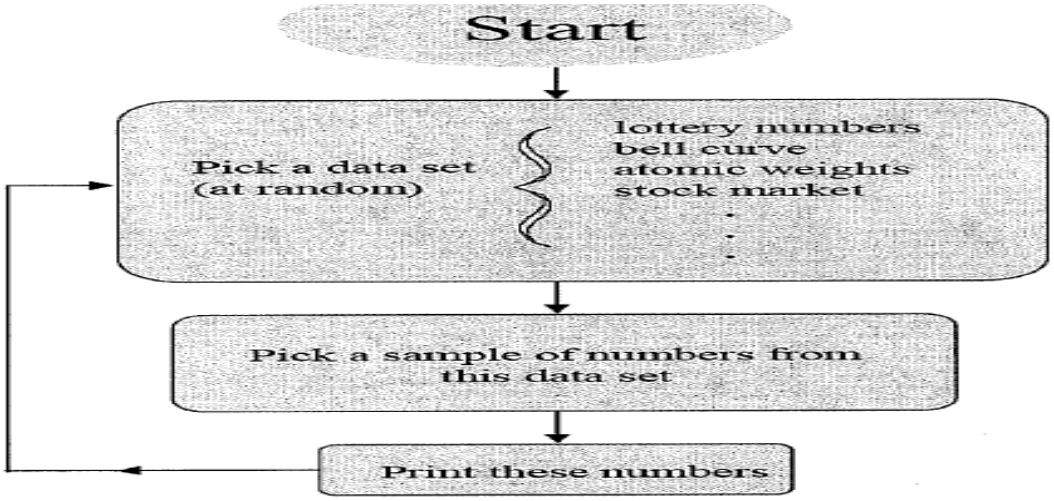
(the number of elements in A that are less than n) / n

The even numbers have natural density 1/2. But the set of integers with leading digit one (as well as the prime numbers with leading digit one) has no natural density. For these sets the ratio oscillates as n increases and never approaches any single value. The key to assigning probabilities to sets of numbers is to develop a measure of density (ie deciding how common events are), so necessary to define the domain of objects appropriately.

So what measure size of these sets? **Average. For set of integers beginning with d the average is $\log_{10}(1+1/d)$**

Hill's theorem: if we repeatedly pick random entries from random distributions, the result tends towards the distribution of Benford's law --that combinations of distributions tend towards the distribution predicted by Benford''s law *even when the original distributions do not* [Hill1996]. This because you are picking across many scales and so this is basically a scale free distribution –> Benford distribution.

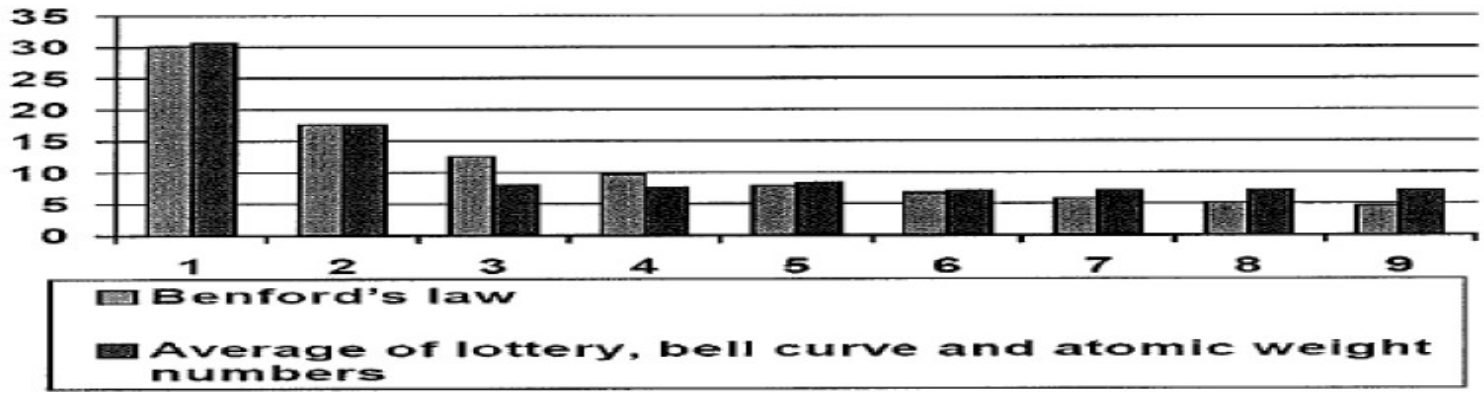
Hill example: you pick lottery numbers from newspaper – uniform; You pick numbers of weights of people – normal distribution; you pick number of automobile accidents – exponential??
Take the % of lead digits of each and average them and you will get Benford!



Log Distribution (Benford's Law)

The "random samples from random distributions" theorem says that if distributions are selected at random (in a neutral way), and samples are taken from each of these distributions, then the resulting data set will have digital frequencies approaching Benford's Law (see Figure 5).

	First Digits								
	1	2	3	4	5	6	7	8	9
Benford's Law	30.1%	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6
Lottery	11%	11	11	11	11	11	11	11	11
Bell Curve	40%	13	8	8	7	7	6	6	5
Atomic Wts.	41%	28	5	4	7	3	4	4	5



Two attributes for Benford: Scale-invariant and Base-invariant

When a distribution is SCALE-INVARIANT, this means we multiply it by some constant λ and the distribution does not change. Write the numbers as $N\ 10^n$ where N is d_1 ... Then $\lambda\ N\ 10^n$ must have the same distribution of the leading digit is $N\ 10^n$. Take log base 10 and do logs, then $\log\ \lambda\ N\ 10^n = \log \lambda + \log N\ 10^n$.

Then you can show that the only distribution that fits this is uniform in log space

SCALE INVARIANCE – the law must be immune to changes in units. P(X) = P(tX) for some change in the units of measurement. Consider a set that fits and see what happens when we change units

stock price $ to Arg Pesos (2.8)	
11	3.1
12	3.4
14	3.9
16	4.5
19	5.3
21	5.9
24	6.7
28	7.8
33	9.2
37	10.4
42	11,8
47	13.2
55	15.4
64	17.9
72	2.02
83	2.3
96	2.7

$Price	Pesos
5 1s	5 1s
3 2s	3 2s
2 3s	3 3s
1 9	1 9

UNIFORM IS NOT SCALE INVARIANT

$s	Euros
1	2
2	4
3	6
4	8
5	10
6	12
7	14
8	16
9	18

BASE INVARIANCE – base 10 is not magical. Should be able to change base and still get the same distribution

BENFORD IS ONLY BASE-INVARIANT DISTRIBUTION

When should it work?

Legally admissible as evidence in criminal cases at the federal, state and local levels. Major accounting firms use computer-assisted audit tools to check data since deviations from Benford's Law should "cause an analyst to question the validity, accuracy, or the completeness of numbers". (Nigrini, Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection. Hoboken, NJ: Wiley, 2012)

ISACA (**Information Systems Audit and Control Association)** tells auditors: "if the audit objective is to detect fraud in the disbursements cycle, the IT auditor **could** use Benford's Law to measure the … leading digits in disbursements compared to the digits' probability. Not work well if sample is small (<100) or subject to special rules such as thresholds and cutoffs. For instance, if a bank's policy is to refer loans at or above US $50,000 to a loan committee, looking just below that approval threshold gives a loan officer the potential to discover loan frauds. " (Simkin, Mark G. "Using Spreadsheets and Benford's Law to Test Accounting Data," ISACA Journal, V1, 2010.

Tracking the Libor Rate (www.escholarship.org/uc/item/2p33x7dk)-- Abrantes-Metz, Villas-Boas, and George " use Benford second digit reference distribution to track the daily London Interbank Offered Rate (Libor) over the period 2005-2008. ...in two recent periods Libor rates depart significantly from the expected Benford.... This raises potential concerns relative to the unbiased nature of the signals coming from the sixteen banks from which the Libor is computed and the usefulness of the Libor as a major economic indicator. integrity of prices." (FYI: Libor is average interest rate calculated by British Banker's Association from interest rates by major London banks. Recently authorities discovered that banks were falsely reporting their rates as to profit from trades, or to give the impression that they were more creditworthy than theywere. Libor underpins approximately $350 trillion in derivatives. Although the FSD's of the Libor interest rates for the time period under study do not span the nine digit space, **the second and following digits data may be** expected to naturally do so.

Scientific fraud in 20 falsified anesthesia papers : (Anaesthesist. 2012 Jun;61(6):543-9) papers known to be falsified by an author were investigated for irregularities with respect to Benford's law using the $\chi(2)$-test and the Z-test. In an analysis of each paper 17 out of 20 studies differed significantly from the expected value for the first digit and 18 out of 20 studies varied significantly from the expected value of the second digit... a meta-analysis using complex mathematical procedures was chosen as a control. The analysis showed a first-digit distribution consistent with the Benford distribution. Thus, the method used in the present study seems to be sensitive for detecting fraud.

Fact and Fiction in EU-Governmental Economic Data German Economic Review Volume 12, Issue 3, pages 243–255, August 2011 To detect manipulations or fraud in accounting data, auditors have successfully used Benford's law as part of their fraud detection processes. Benford's law proposes a distribution for first digits of numbers in naturally occurring data. Government accounting and statistics are similar in nature to financial accounting. In the European Union (EU), there is pressure to comply with the Stability and Growth Pact criteria. Therefore, like firms, governments might try to make their economic situation seem better. In this paper, we use a Benford test to investigate the quality of macroeconomic data relevant to the deficit criteria reported to Eurostat by the EU member states. **We find that the data reported by Greece shows the greatest deviation from Benford's law among all euro states.**

Benford's law and Theil transform of financial data P. Clippe and M. Ausloos (marcel.ausloos@ulg.ac.be) Physica A: Statistical Mechanics and its Applications, 2012, vol. 391, issue 24, pages 6556-6567-- the present paper concerns the Antoinist community financial reports–a community which appeared at the end of the 19th century in Belgium.... there is common suspicion about sect finances. In that spirit, the Antoinist community yearly financial reports, income and expenses, are hereby examined through ...Benford's law, (which) is often used as a test about possible accounting wrongdoings. On the other hand, Benford's law is known to be invariant under scale and base transformation. Therefore, as a further test, of both such data and the use of Benford's law, the yearly financial reports are nonlinearly remapped through a sort of Theil transformation …

Mir, T. A. (2012). The law of the leading digits and the world religions. Physica A: Statistical Mechanics and its Applications, 391(3), 792–798.

1. Abstract

Benford's law states that the occurrence of significant digits in many data sets is not uniform but tends to follow a logarithmic distribution such that the smaller digits appear as first significant digits more frequently than the larger ones. We investigate here numerical data on the country-wise adherent distribution of seven major world religions i.e. Christianity, Islam, Buddhism, Hinduism, Sikhism, Judaism and Baha'ism to see if the proportion of the leading digits occurring in the distribution conforms to Benford's law. We find that the adherent data of all the religions, except Christianity, excellently does conform to Benford's law. Furthermore, unlike the adherent data on Christianity, the significant digit distribution of the three major Christian denominations i.e. Catholicism, Protestantism and Orthodoxy obeys the law. Thus in spite of their complexity general laws can be established for the evolution of the religious groups.

Table 3: The significant digit distribution of country-wise adherents of all religions

First Digit	1	2	3	4	5	6	7	8	9	Total
N_{Obs}	289	169	118	84	66	69	64	36	49	**944**
N_{Ben}	284.2	166.2	117.9	91.5	74.7	63.2	54.7	48.3	43.2	
Error	14.1	11.7	10.2	9.1	8.3	7.7	7.2	6.8	6.4	

Table 1: The significant digit distribution of country-wise Christian, Catholic, Protestant, Orthodox and Muslim populations

First Digit	Christian (205)	Catholic (197)	Protestant (171)	Orthodox (42)	Muslim (184)
1	50 (61.7±6.6)	47 (59.3±6.4)	48 (51.5±6.0)	12 (12.6±3.0)	60 (55.4±6.2)
2	30 (36.1±5.4)	34 (34.7±5.3)	25 (30.1±5.0)	5 (7.3±2.5)	43 (32.4±5.2)
3	29 (25.6±4.7)	31 (24.6±4.6)	26 (21.4±4.3)	7 (5.2±2.1)	21 (23.0±4.5)
4	27 (19.9±4.2)	18 (19.1±4.1)	15 (16.6±3.9)	3 (4.1±2.0)	13 (17.8±4.0)
5	14 (16.2±3.9)	19 (15.6±3.8)	18 (13.5±3.5)	4 (3.3±1.7)	7 (14.6±3.7)
6	15 (13.7±3.6)	10 (13.2±3.5)	10 (11.4±3.3)	4 (2.8±1.6)	11 (12.3±3.4)
7	18 (11.8±3.3)	10 (11.4±3.3)	9 (9.9±3.0)	5 (2.4±1.5)	13 (10.7±3.2)
8	6 (10.5±3.1)	15 (10.1±3.1)	12 (8.7±2.9)	2 (2.1±1.4)	10 (9.4±3.0)
9	16 (9.4±3.0)	13 (9.0±2.9)	8 (7.8±2.7)	0 (1.9±1.3)	6 (8.4±2.8)
Pearson χ^2	**16.419**	**10.143**	**5.208**	**6.946**	**10.646**

Table 2: The significant digit distribution of country-wise Baha'i, Buddhist, Hindu, Sikh and Jewish populations

First Digit	Baha'i (175)	Buddhist (129)	Hindu (97)	Sikh (47)	Jew (107)
1	61 (52.9±6.1)	38 (38.8±5.2)	26 (29.2±4.5)	12 (14.1±3.1)	42 (32.2±4.7)
2	35 (30.8±5.0)	21 (22.7±4.3)	16 (17.1±3.7)	11 (8.3±2.6)	13 (18.8±4.0)
3	20 (21.9±4.4)	15 (16.1±3.7)	12 (12.1±3.2)	6 (5.9±2.3)	15 (13.4±3.4)
4	11 (16.9±3.9)	15 (12.5±3.4)	10 (9.4±2.9)	2 (4.5±2.0)	6 (10.4±3.1)
5	10 (13.8±3.6)	10 (10.2±3.1)	8 (7.7±2.6)	5 (3.7±1.8)	12 (8.5±2.8)
6	16 (11.7±3.3)	8 (8.6±2.8)	11 (6.5±2.5)	1 (3.1±1.7)	7 (7.2±2.6)
7	9 (10.1±3.1)	10 (7.5±2.6)	3 (5.6±2.3)	5 (2.7±1.6)	6 (6.2±2.4)
8	5 (8.9±2.9)	7 (6.6±2.5)	3 (5.0±2.1)	2 (2.4±1.5)	3 (5.5±2.3)
9	8 (8.0±2.8)	5 (6.0±2.4)	8 (4.4±2.0)	3 (2.1±1.4)	3 (4.9±2.2)
Pearson χ^2	**8.648**	**1.786**	**8.45**	**6.863**	**10.158**