Ciencias Económicas 30-No.2: 2012 / 179-200 / ISSN: 0252-9521

STATISTICIAN'S NEW ROLE AS A DETECTIVE – TESTING DATA FOR FRAUD

Alex Ely Kossovsky[1]

RESUMEN

El objetivo de este trabajo es proporcionar al estadístico con un método para la difícil tarea de decidir si un determinado conjunto de datos podría haber sido inventado en forma fraudulenta o en aparente autenticidad. Esto no se hace mediante el examen de los números en sí, pero, sorprendentemente, se hace investigando el lenguaje digital utilizado en escribir esos números! Lo que las letras son para las palabras, los dígitos son para los números. La técnica se basa en la Ley de Benford, una ley estadística que se refiere a la consistencia y predictibilidad de las proporciones relativas a los dígitos que ocurren en los datos típicos de la vida real, estableciendo que dígitos bajos son mucho más frecuentes que los dígitos altos. La ley es extremadamente útil como una herramienta para detectar el fraude, especialmente el fraude fiscal, ya que los estafadores tienden a inventar números donde los dígitos tienen aproximadamente las mismas proporciones, debido a la intuición equivocada de que todos los dígitos aparecen en los datos con igualdad de oportunidades. Al comparar la distribución teórica de Benford con la distribución real de dígitos de los datos de la contabilidad de las empresas, el estadístico puede fácilmente descubrir el fraude en relación con datos falsos o inventados. Estas pruebas forenses digitales son ahora los procedimientos habituales en los departamentos de los ingresos fiscales de la mayoría de los gobiernos de todo el mundo, así como en la contabilidad de las grandes compañías de auditoría.

PALABRAS CLAVES: DETECCIÓN DE FRAUDE FISCAL, LENGUAJE DIGITAL, ANÁLISIS FORENSE DE DATOS, TERREMOTO, COMPUESTOS QUÍMICOS, DATOS ASTRONÓMICOS.

ABSTRACT

The objective of this paper is to provide the statistician with a method for the challenging task of deciding whether a given data set might have been invented in a fraudulent way or appearing authentic. This is done not by examining the numbers themselves, but surprisingly, rather by investigating the digital language utilized in writing those numbers! What letters are to words, digits are to numbers. The technique relies on Benford's Law, a statistical law referring to the consistent and predictable relative proportions of digits occurring in typical real-life data, stating that low digits are much more frequent than high digits. The law is immensely useful as a tool to detect fraud,

1 El autor es graduado del City University of New York en Estadística y Matemática Aplicada, y ha realizado investigaciones en Benford's Law durante los últimos años. Correo electrónico: akossovs@yahoo.com.

especially tax fraud, since cheaters inventing fake data mistakenly write them with all digits having about the same proportion due to the erroneous intuition that all digits come with equal chances. By comparing theoretical Benford digit distribution to the actual digit distribution within the accounting data provided by companies, the statistician can easily discover fraud relating to fake and invented data. These digital forensic tests are now standard procedures in most of the Tax Revenue Departments of governments worldwide, as well as in large accounting and auditing companies.

KEY WORDS: TAX FRAUD DETECTION, DIGITAL LANGUAGE, FORENSIC DATA ANALYSIS, EARTHQUAKE, CHEMICAL COMPOUNDS, ASTRONOMICAL DATA

I. INTRODUCTION

Recent statistical discoveries allow the statistician to utilize known digital patterns in typical data to detect fraud. Previously, the task of the statistician was to analyze and summarize data, show patterns, and make predictions, but never to decide on the authenticity of the provided data itself. Data provided to the statistician was traditionally always taken as a given without an ability to authenticate. On the other hand, there is always a very strong need on the part of Tax Authorities worldwide and accounting companies to obtain professional statistical advice as to how to detect fake data. Data could be faked to reduce and to under-report revenues in order to pay less tax, as well as to inflate revenues at times in order to impress investors and present the company as being financially healthy. The enormous amount of tax money saved per year for various governments worldwide regularly via forensic digital analysis utilizing Benford's Law can't be underestimated, it is huge. Similar benefits in savings and discontinuation of on-going fraud schemes within companies by insiders, fraudulent treasurers, and financial officers are also extremely valuable, and are achieved by the same digital techniques.

The technique relies on Benford's Law, a statistical law referring to the consistent and predictable relative proportions of digits occurring in typical real-life data, stating that low digits such as 1, 2, and 3 are much more frequent than high ones such as 7, 8, and 9. For example, numbers whose first digit on the left is 1 are very common, occurring in about 30.1% of values, while numbers beginning with digit 9 are relatively rare, occurring only about 4.6% of values. The main reason that this forensic test is at all possible springs from the fact that cheaters inventing fake data almost always erroneously write them with all digits having about the same proportion due to the mistaken intuition that all digits have equal chance of occurrence. By comparing theoretical Benford digit distribution to the actual digit distribution within the accounting data provided by companies, the statistician can decide if data appear suspicious or normal. During the past 15 years most Tax Revenue Departments of governments worldwide as well as large accounting and auditing companies have adopted these digital forensic tests as their standard procedures, and run them on a regular basis. The results of this revolutionary new technique has been a great increase in the revenue of tax collection money, as well as numerous cases of fraud that have been detected, stopped, and prevented from further exploiting or ruining financially healthy companies.

II. THE LEADING DIGITS PHENOMENA

Leading digits (LD) or first significant digits are the first digits of numbers appearing on the left. Such a digit is called "the leader" of the number because all other digits follow it. For 567.34

the leading digit is 5. For 0.0367 the leading digit is 3, as we discard the zeros. For the lone integer 6 the leading digit is 6. For negative numbers we simply discard the sign.

613 ------> digit 6
0.0002867 ------> digit 2
7 ------> digit 7
-7 ------> digit 7
1,653,832 ------> digit 1
-0.456398 ------> digit 4

The temptation here (even for the statistician!) is to believe one's own intuition that for numbers occurring in everyday typical situations, all digits should have an equal chance of occurring, and thus uniformly distributed. But let's look at some surprising results from the closing prices and daily volume of stocks traded on The New York Stock Exchange (Bolsa) on December 23, 2011. We choose the first 31 companies on top of the alphabetically-sorted list as our random small sample:

FIGURE 1
PRICES AND VOLUME OF STOCKS TRADED ON THE NEW YORK STOCK EXCHANGE

Stock Symbol	Closing Price	Volume
A	$ 30.74	1,124,700
AA	$ 38.32	5,950,900
AAI	$ 7.03	533,700
AAP	$ 34.09	430,100
AAR	$ 22.14	8,600
AAV	$ 11.01	263,800
AB	$ 60.86	335,400
ABA	$ 25.75	4,000
ABB	$ 25.12	2,627,700
ABC	$ 41.48	478,600
ABD	$ 14.03	264,200
ABG	$ 14.24	164,500
ABH	$ 9.68	992,700
ABI	$ 34.42	791,000
ABK	$ 9.94	1,688,700
ABM	$ 19.88	140,100
ABN	$ 57.62	29,500
ABN PRE	$ 21.49	28,000
ABN PRF	$ 23.58	5,800
ABN PRG	$ 22.15	46,100
ABR	$ 15.92	254,700
ABT	$ 53.23	2,336,000
ABV	$ 85.17	406,200
ABV C	$ 77.19	5,400
ABW PRA	$ 25.02	1,900
ABX	$ 53.55	2,574,500
ACC	$ 26.52	147,300
ACE	$ 55.09	1,216,700
ACE PRC	$ 24.92	11,300
ACF	$ 14.50	597,600
ACG	$ 8.39	193,300

Source: https://nyse.nyx.com/

About half of the numbers here start with digit 1 or digit 2! Here is the exact LD distribution for this limited set of 31 companies above. It should be noted that almost all other such subsets down the long list on the NYSE website yield quite similar results:

FIGURE 2
LEADING DIGITS OF STOCK PRICES AND VOLUME

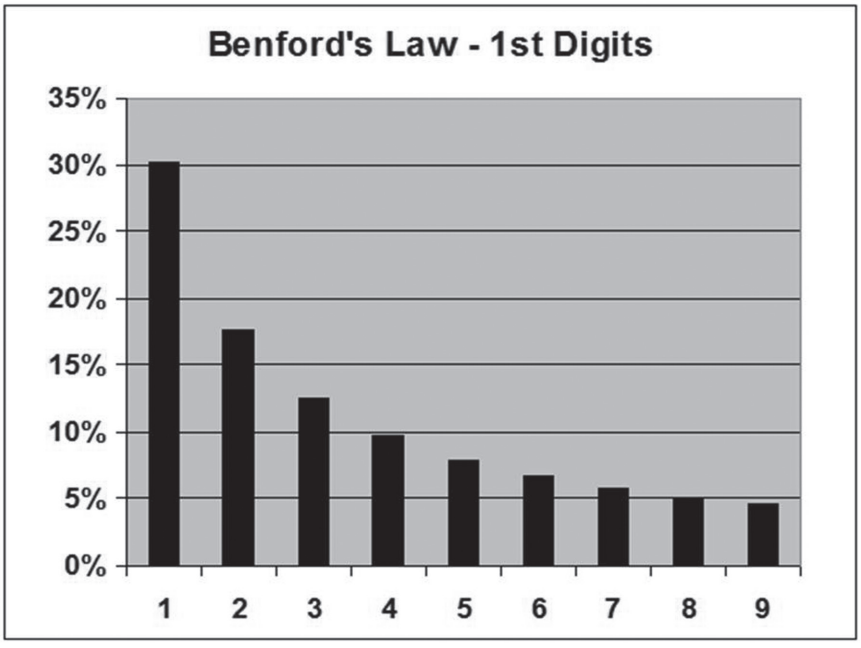
Digit	Price	Volume
1	19.4%	29.0%
2	29.0%	25.8%
3	12.9%	3.2%
4	3.2%	16.1%
5	12.9%	16.1%
6	3.2%	0.0%
7	6.5%	3.2%
8	6.5%	3.2%
9	6.5%	3.2%

Source:: UCR- Calculaciones del Bursa – New York Stock Exchange

Simon Newcomb in 1881 and then Frank Benford in 1938 discovered that low digits lead much more often than high digits in everyday and scientific data and arrived at the exact expression of Probability[1st digit is d] = $LOG_{10}(1+1/d)$ being the probability that digit d is leading. This set of proportions is known as the logarithmic distribution, and the law is known as Benford's Law. For example, P[1st digit is 1] = $LOG_{10}(1+1/1) = LOG_{10}(2) = 0.301$.

B.L. (1st Digits) = {30.1%, 17.6%, 12.5%, 9.7%, 7.9%, 6.7%, 5.8%, 5.1%, 4.6%}.

FIGURE 3
CHART OF BENFORD'S LAW - 1ST DIGITS



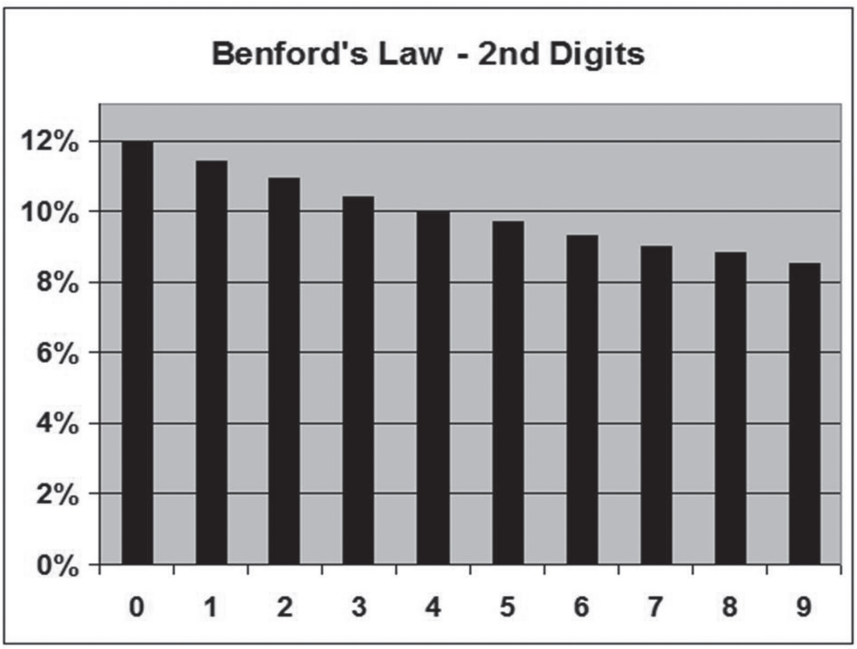
Source: http://en.wikipedia.org/wiki/Benford's_law
http://www.auditnet.org/articles/JFA-V-1-17-34.pdf

The law also describes an exact distribution for the second order digits, but here proportions among digits are more equal. For example, the 2nd leading digit (from the left) of 603 is digit 0, of 0.0002867 it's digit 8, and of 1,653,832 it's digit 6. It is noted that for the 2nd and higher orders, digit 0 is also included, whereas for the 1st digit order it is excluded. The exact 2nd order distribution for all 10 digits according to Benford's Law is:

B.L. (2nd Digits) = {12.0%, 11.4%, 10.9%, 10.4%, 10.0%, 9.7%, 9.3%, 9.0%, 8.8%, 8.5%}.

613 ------> digit 1
0.0002867 ------> digit 8
1,653,832 ------> digit 6
-0.456398 ------> digit 5
603 ------> digit 0

FIGURE 4
CHART OF BENFORD'S LAW – 2ND DIGITS



Source: http://www.auditnet.org/articles/JFA-V-1-17-34.pdf

Digital proportion for the 2nd order is not nearly as skewed in favor of low digits as is the case for the 1st order. The 3rd order digit distribution is even more equal than the 2nd order. And finally there is almost total digital equality for the 4th and higher orders.

The probability of any First-Two-Digit combination (called FTD), such as 34, and exemplified in numbers such as 3487, 0.0341, 340 etc. is given by the formula:

Probability[1st digit is p AND 2nd digit is q] = $LOG_{10}(1 + 1/pq)$.

For example, P(10) = LOG(1 + 1/10) = LOG(1.1) = 0.0414

P(25) = LOG(1+1/25) = 0.0170 and P(99) = LOG(1+1/99) = 0.0044.

When numbers in the data set are long enough (i.e. many digits per number), and typically this is the case, namely that we almost always have plenty of digits in each value (say over 4 or 5 digits), then the last digit distribution (digits on the right-most side) should be about uniform with

equal probability of 1/10 for each digit. Also, the last-two digit combinations (on the right) should also show uniformity with an equal probability of 1/100. This is so since there are 100 possibilities here, namely {00, 01, 02, …, 97, 98, 99}.

As mentioned by Hill (1995) in the first section titled "The Significant-Digit Law", as well as in the 2nd section "Empirical Evidence", the validity of Benford's Law has been observed and verified in numerous domains including finance, accounting, economics, census data, physics, astronomy, chemistry, and geology, to mention just a few.

According to Durtschi C. et al. (2004), Hill (1995), and others in the literature, the following short summery about applicability can be given. The specific types of data that ARE Benford include: any well-mixed data from a variety of sources, almost all accounting data such as account receivable, account payable, revenues, expenses, election results by city/town or province (if free and fair, not manipulated), population and other census data, size of files in megabytes on any typical large computer, sport data, the list of all the physical constants in Physics and Chemistry (combined), house numbers in address data, data derived from multiplication processes such as exponential growth and decay, the Fibonacci series. While the specific types of data that are NOT Benford include: phone/cell numbers, lottery numbers, code and index numbers, serial numbers, purchase ID order numbers, ATM withdrawal amounts, any other data with pre-assigned values, data with artificial built-in minimum or maximum values.

Benford's Law is a mathematical and statistical fact about how numbers are USED and OCCUR in everyday typical situations, expressing physical quantities as well as abstract entities we wish to record. But the law is NOT a mathematical law purely about our number system itself - totally divorced from its use. It is indeed also a physical and scientific law regarding quantitative measurements, and its scope covers all disciplines, as it is being almost universally observed. Rarely do we encounter such a prevalent law or regularity spanning all disciplines in science, linking and connecting various fields of study, from physics, chemistry, astronomy, and geology, to economics, finance, accounting, and so forth.

III. THE LOGARITHMIC AS REPEATED MULTIPLICATIONS

Repeated multiplication processes effectively drive numbers toward the logarithmic distribution in the limit. This is so only if all intermediate values are considered and retained as part of that long sequence of numbers. Classical examples include: (I) Money invested in a bank account and locked in for 30 years at 5% interest rate, while yearly snapshots of account balance are taken each December 31. (II) Bacteria in the lab tube growing at 20% per hour, while count is conducted on an hourly basis. Both are examples of exponential growth having a fixed factor. Note the **deterministic** nature of this process, absence any random factor involved, as oppose to statistical and typical real life data which are intrinsically **random**, and in Benford's Law this distinction is crucial and **applicable.**

We consider the arbitrarily chosen value of 8 being repeatedly multiplied by another arbitrary number 3 (exponential growth in essence). This yields the following table:

FIGURE 5
SIMPLE MUTLIPLICATION PROCESS, $8*3^I$ AND ITS LD DISTRIBUTION

$8*3^i$	$8*3^i$ series	1st digit
$8*3^0$	8	8
$8*3^1$	24	2
$8*3^2$	72	7
$8*3^3$	216	2
$8*3^4$	648	6
$8*3^5$	1944	1
$8*3^6$	5832	5
$8*3^7$	17496	1
$8*3^8$	52488	5
$8*3^9$	157464	1
$8*3^{10}$	472392	4
$8*3^{11}$	1417176	1
$8*3^{12}$	4251528	4
$8*3^{13}$	12754584	1

Digit	Proportion
1	35.7%
2	14.3%
3	0.0%
4	14.3%
5	14.3%
6	7.1%
7	7.1%
8	7.1%
9	0.0%

Source: UCR -Calculaciones del - multiplication 8 repeat times 3

Other arbitrarily chosen values give similar results, especially if extended much farther than the few 14 sequences in the above table. Hence, the above result is quite general and representative, and in the limit the logarithmic emerges.

Geometric series and exponential growth or decay exhibit logarithmic behavior, and that can be said for almost any growth rate (any factor) and any starting value, yielding the logarithmic almost exactly given that enough elements are being considered.

The Fibonacci series {1, 1, 2, 3, 5, 8, 5+8, etc.} where {1, 1} are arbitrarily chosen, and subsequent elements being the addition of the previous (last) two elements, approaches approximately a repeated multiplication process, with the golden ratio 1.618 being the factor, explaining its almost perfect logarithmic behavior. Further readings on the Fibonacci series and the golden ratio can be found in Burton D. (1993) "The History of Mathematics".

IV. HILL'S SUPER DISTRIBUTION

After many decades since the re-discovery of the phenomena in 1938 by Frank Benford, and after numerous failed attempts at proofs, a rigorous mathematical explanation was given in 1995 by Theodore P. Hill, demonstrating that a super distribution consisting of infinitely many distributions, all mixed together and defined on the positive x-axis, is logarithmic in the limit as the number of such distributions approaches infinity. For practical consideration, it's enough to have just a few mixed densities to observe 'almost Benford' behavior, and therefore its application is widespread and typical, although surely many real-life data types, such as those pertaining to single-issue

observations, and many others, can NOT be modeled on Hill's construct at all. Hill's proof explains and covers only a part of the entire phenomena. Hill's argument then is that those relevant everyday data types pertaining to his mathematical model are simply compositions of a variety of variables and distributions, and that picking numbers from such real life data is schematically equivalent to picking from his abstract random number arising from a distribution of distributions. The most remarkable feature in Hill's distribution of distributions is the multiple levels of randomness involved. One has to first randomly select a distribution type, then continue to select values for its parameters, and finally to pick a particular number from that random distribution, all in all involving 3 distinct levels of randomness.

V. THE SCALE INVARIANCE PRINCIPLE

What happens to digits for data in Costa Rica on gasoline purchases in Colones for example when converted into US Dollars or Euros? And what happens to digit distribution of data on weights of people in Costa Rica in Kilograms (40-90 kilos most typical), when translated into British Pounds (80-180 lbs most typical)? Surely digital distributions in the two cases above change dramatically after such a scale conversion! Does Benford's Law then depend on our specific global units and scales system we use at this current epoch? Would a sudden global change in units affect the law? Surprisingly, the answer is a definite and decisive NO! Surely, LD of individual types of data sets would be dramatically affected by such unit and scale change, yet, a large combination of data types all mixed in as in Hill's model would not be affected in the least, because of trades offs and offsetting changes acting in opposite directions leading to cancellations and leaving it unaltered at the end. Hence the law is totally independent of units and scales. Moreover, any single-issue data type X_i that is logarithmic in its own right, is also so after multiplicative conversion by ANY factor. That is, the newly transformed $F*X_i$ data set is logarithmic just the same.

A novel approach to leading digits is one based on the scale invariance principle. Pinkham (1961) employed the scale invariance argument to claim that if there is indeed any universal law for significant digits, then it should be independent of the units and scales used by society, because scales are cultural, arbitrary, and do not represent any fundamental properties of numbers or nature. In other words, if Benford's Law were to be dependent on humanity using the Kilometer as oppose to the Mile, or the Kilogram as oppose to the Pound, then the law would not be universal! Pinkham (1961) then further demonstrated mathematically that the logarithmic distribution is scale-invariant, and that it's the only distribution with such a property; therefore any first digit law that is independent of choices of scale must be the logarithmic!

VI. AVERAGING SCHEME AS A MODEL FOR TYPICAL DATA

In another quite different approach to Benford's Law, Kossovsky (2006) has attempted to construct a limited but well-structured mathematical model that is approximately capable of capturing or representing many cases of typical usage of numbers, and thus to arrive at another explanation for this digital phenomenon. Real life data sets typically start on the left at 0 or 1, or some other very low value, but usually end much higher to the right at varying upper limits depending on topic and type. And so we consider what would happen (collectively) to leading digits distributions of numerous intervals (made only of the integers), short and long, and those intermediate ones in between, all similarly starting at 1, called Lower Bound, or LB, while differing strongly in their lengths, namely the Upper Bound, called UB. The plan is then to obtain an aggregate LD distribution representing all the intervals simply by taking the average of the LD distributions of all the individual intervals. One such scheme, with LB=1, and with UB varying from 10 to 100 in steps of 1 unit (integer) yielded the average of: {24.5%, 18.4%, 14.5%, 11.7%, 9.4%, 7.6%, 6.0%, 4.6%, 3.3%}.

This only resembles the logarithmic, but it is not quite close enough. What went wrong? An argument can be raised that the scheme is not complex enough, and that it's too rigid, namely that it doesn't average enough things as it pertains only to such a structure with BU varying between 10 and 100 - two arbitrarily chosen values. The next (more complex) level then, is to let the edges of UB themselves vary progressively, 1 integer at a time, from UB_minimum to UB_maximum, making the selection of UB variation less arbitrary, while lower bound is still fixed at 1. In other words, to average results from multiple such averaging schemes all with different arrangements of UB layouts, so as to give a more general result. This more complex scheme gets much closer to the logarithmic, but it is still not close enough. The next higher order scheme is to vary UB_maximum itself also from UB_maximum_LOW to UB_maximum_HIGH, which then gives results that are even closer to the correct logarithmic distribution. Conclusion: What we really need here is an infinite scheme! Flehinger (1966) has given a rigorous mathematical proof that such an infinite algorithm would converge to the logarithmic in the limit. Many real life data types can be directly model on such a scheme (most typically the house number in address data), and so their logarithmic behavior can be satisfactory explained by such averaging algorithms.

VII. CHAINS OF DISTRIBUTIONS

An alternative point of view of the averaging schemes, putting them in a statistical framework, has been suggested by Kossovsky (2006). The idea is to consider those varying upper bounds as originating from the random process of the Uniform distribution. In other words, the corresponding view here is to examine leading digits of random numbers from the continuous uniform on (0, B) standing for the generic single interval, and where parameter B, the upper bound, is itself a random number drawn from another uniform distribution on (0, C). Schematically this is written as: Uniform(0, Uniform(0, C)). Conveniently, the ranges for the chains here start from 0 as oppose to the usage of 1 as LB earlier for the averaging schemes. Kossovsky argued that these types of chains of distributions are essentially mirror images of the averaging schemes, and gave numerous digital results regarding a large variety of chains of distributions employing many classical distributions all tied in sequentially in many different styles of dependencies. Traditionally parameters of distributions have always been thought of exclusively as constants, fixed by the particular nature of the data on hand, yet this new unorthodox approach leads to better understanding of Benford's Law and the digital phenomena in general, and it initiates the study of the behavior of such complex statistical constructs which might have applications in other contexts and disciplines in the future. Attempts here to perform simulations of the model Uniform(0, Uniform(0, C)) resulted in various LD distributions, somewhat close to the logarithmic, and depending on the particular value of C, as might be expected.

The next natural step is U(0, U(0, U(0, K))). This yielded better conformity with the logarithmic. Results for the next logical step, namely U(0, U(0, U(0, U(0,M)))) are:

{30.1%, 17.6%, 12.5%, 9.7%, 7.9%, 6.7%, 5.8%, 5.1%, 4.6%} which is almost Benford.

Clearly, what we need here is an infinite chain to obtain the logarithmic! Therefore, picking a number from a truly random interval, without even specifying directly parameters leads to the logarithmic distribution.

Kossovsky conjectured that this result is much more general, pertaining to most other distributions forms, and not only to the Uniform, even including combinations of a variety of mixed forms. Convergence here is actually very rapid, and there is 'no need to go to infinity' to obtain digit distribution that is approximately very close to Benford. Here is one notable demonstration using the notation Normal(mean, s.d.):

Normal(*Uniform*(0, *chi-sqr*(a die of 6 sides)), *Uniform*(0, 2))

Resulting in: {30.1%, 17.6%, 12.5%, 9.7%, 7.9%, 6.7%, 5.8%, 5.1%, 4.6%}.

Another important result in general here is that a 2-sequence chain in which all parameters are derived from logarithmic distributions is logarithmic immediately without any need to expand infinitely. In symbols: Any Distribution(Any Benford) is Benford.

Miller (2008) gave rigorous mathematical proofs to some of Kossovsky's conjectures, including the result of the 2-sequence chain based on logarithmic parameter.

VIII. NATURE'S WAY OF COUNTING SINGLE-ISSUE PHENOMENON

Another common source of the logarithmic distribution and one profound reason for its prevalence in real life data is its physical manifestation. Data sets of single-issue quantities, such as amount of water in river flow, earthquake depth, time between successive earthquakes, rotation rates of spinning star remnants known as pulsars, population data, as well as countless others pieces of specific physical data sets, are logarithmic in their own right, individually considered, without having to be schematically averaged out or get mixed with anything else.

All this represents something quite new and striking in Benford's Law, requiring a radically different explanation than that supplied by Hill's findings, or by Flehinger's averaging schemes and Kossovsky's chains above, where data was found be to logarithmic following our man-made and artificial aggregation, compilation, and mixing.

The evidence that Benford's Law is a common feature across the physical sciences spanning every discipline is quite compelling, and more so with the recent avalanche of additional findings and testing. Empirically, the logarithmic is found not only in data sets rooting in macrocosmic systems (stars/galaxies/rivers), but also in microcosmic systems (atomic/subatomic particles/molecules). Moreover, Benfrod's Law is observed in Dynamic systems (earthquakes/active) as well as in Static systems (chemical molar mass/passive). Statisticians and mathematicians will continue to debate the theoretical justification for Benford's Law existence, and the fact that it pops up so frequently in numerous natural phenomena wouldn't surprise them in the least, yet it does often shock many scientists!

Data on all known exoplanets in our Milky Way Galaxy, as of early September 2012, containing 834 planets outside the solar system, is one remarkable example of the prevalence of Benford's Law in the physical world. The data includes values for planets' mass, angular distance, semi-major axis size, orbital eccentricity, and orbital period. The very fact that 5 different aspects or measurements of a single physical reality are all nearly Benford is quite intriguing! The table in figure 6 shows the first digit distributions of these 5 variables pertaining to the same physical set of 834 planets. Results are quite close to the logarithmic, and deviations are not extreme, in spite of the fact that this data set is (statistics-wise) very small, representing a tiny portion of the estimated 160 billion or so star-bound planets that exist in our Galaxy. It is noted that only 30 planets were discovered in the years 1989 to 1999, while the vast majority of them, 804 in all, are very recent discoveries during the years 2000 to 2012 – namely during the first 12 years of this new millennium.

FIGURE 6
FIVE DIFFERENT ASPECTS OF EXOPLANETS ARE ALL BENFORD!

DIGIT	Planet's mass	Angular distance	Semi-major axis size	Orbital eccentricity	Orbital period
1	30.0%	30.0%	28.2%	28.2%	27.4%
2	18.8%	16.6%	19.4%	21.7%	14.6%
3	11.9%	13.7%	11.9%	13.6%	17.7%
4	7.4%	8.0%	13.4%	11.6%	13.9%
5	8.0%	7.7%	9.2%	7.2%	7.8%
6	7.6%	8.6%	5.7%	6.1%	5.6%
7	7.4%	5.1%	4.1%	4.2%	4.5%
8	4.6%	5.4%	5.0%	4.6%	4.2%
9	4.4%	5.0%	3.1%	2.8%	4.5%

Fuente: http://exoplanet.eu/catalog/

The tables shown in Figures 7 and 8 are courtesy of the geologist Malcolm Sambridge of the Australian National University in Canberra. The data was complied by him and his colleagues, providing a list of natural phenomena with properties that follow Benford's law. In order presented in the tables, it includes: (1) the depths of almost 250,000 earthquakes that occurred worldwide between 1989 and 2009, (2) the time interval in seconds between consecutive earthquakes worldwide in the period 01/01/1970 to 12/31/2009 with no restrictions on geographical position, depth or magnitude, (3) the rotation rates or frequencies of spinning remnants of dead stars also known as pulsars, given in Hz, from the ATNF catalogue, (4) river lengths in Canada, (5) global monthly averaged temperature anomalies from the gistemp database over the period 1880-2008 measured in degrees, (6) total numbers of cases of 18 infectious diseases within countries reported to the World Health Organization by 193 countries worldwide in 2007, (7) the Earth's geomagnetic field model gufm1, (8) time in years between the 93 particular reversals of the Earth's Geomagnetic field, (9) regional body wave seismic model, (10) the global seismic tomography shear wavespeed model of the Earth mantle, (11) the anisotropic shear wave mantle model saw642an, (12) the brightness of gamma rays that reach Earth as recorded by the Fermi Gamma-ray Space Telescope across the galactic in first 11 months of operation.

FIGURE 7
EARTHQUAKE, PULSAR, RIVER, TEMPERATURE, AND DISEASE DATA ARE BENFORD

Digit	Earthquake depths	Time between earthquakes	Pulsars rotation frequencies	River lengths (Canada)	Global Temperature anomalies	Global Infectious disease
1	31.6%	29.1%	33.9%	21.5%	27.7%	33.7%
2	16.9%	17.2%	20.7%	17.1%	19.4%	16.7%
3	14.0%	12.6%	12.7%	14.6%	12.7%	13.2%
4	8.7%	10.0%	7.6%	15.8%	12.1%	10.7%
5	7.0%	8.2%	5.3%	9.5%	8.9%	7.3%
6	7.4%	6.9%	5.0%	6.3%	5.4%	5.4%
7	5.3%	5.9%	4.9%	6.3%	6.6%	4.6%
8	4.6%	5.2%	4.7%	5.1%	4.3%	5.1%
9	4.4%	4.6%	4.9%	3.8%	2.8%	3.3%
Data points	248,915	2,258,653	1,861	158	1,527	987

Source: http://www.asiapacific-mathnews.com/01/0104/0001_0006.pdf
http://rses.anu.edu.au/~malcolm/papers/pdf/Sambridge-etal2011APMN.pdf
https://researchers.anu.edu.au/researchers/sambridge-ms

FIGURE 8
GEOLOGICAL AND CELESTIAL DATA THAT ARE BENFORD

DIGIT	Geomag Field	Geomagnetic reversals	Seismic P wavespeed (SW-Pacific)	Whole mantle wavespeed	Whole Earth shear anisotropy	Fermi telescope ray fluxes
1	28.9%	32.3%	30.0%	32.0%	30.7%	30.3%
2	17.7%	19.4%	17.6%	17.4%	14.6%	17.9%
3	13.3%	13.9%	13.3%	12.4%	11.0%	13.0%
4	9.4%	11.8%	9.8%	9.1%	9.3%	9.9%
5	8.1%	5.3%	7.9%	7.3%	8.6%	7.6%
6	6.9%	4.3%	6.4%	6.5%	7.6%	7.0%
7	6.1%	3.2%	5.6%	5.9%	6.9%	5.2%
8	5.1%	5.4%	4.9%	4.9%	6.0%	5.2%
9	4.5%	4.3%	4.5%	4.6%	5.3%	2.7%
Data points	36,512	93	423,776	10000	20,544	1451

Source: http://www.asiapacific-mathnews.com/01/0104/0001_0006.pdf
http://rses.anu.edu.au/~malcolm/papers/pdf/Sambridge-etal2011APMN.pdf
https://researchers.anu.edu.au/researchers/sambridge-ms

Besides measuring earthquake depths, Sambridge's team also examined vertical displacements of the ground in Peru as the tsunami-triggering Sumatra-Andaman earthquake of 2004 progressed. A set of ground shifts before the earthquake proper did not follow Benford's law, but shifts that occurred during the quake itself did. The team also examined seismic data recorded at the same time by a station in Canberra. The overall patterns in the shifts persisted but the exact extent of the adherence to Benford's Law varied differently over time than in the Peruvian measurements. The team then looked more closely at Canberra seismograms and found that they were consistent with a minor, local earthquake occurring at the same time, which could be the source of the discrepancy between the two measurements. "That's the first time I know of where something physical like that was actually discovered using Benford's law," said Ted Hill upon learning of the team's work. Detailed account of the team's methods and conclusions can be found in Sambridge's article titled "Benford's Law in the Natural Sciences". The significance of Sambridge's work has been the demonstrated ability to detect an earthquake just from the first digit distribution of the seismic waveforms data, and in spite of the apparent loss of the complex information contained in the actual data itself –being reduced down to just its first digit proportions! Inspired by this remarkable example, quantum physicists are recently applying Benford's Law to detect quantum phase transitions with success. In general, this shows that the examination of leading digit distribution has the potential to detect changes in the physical world. Sambridge suggestion for application is to first determine empirically which phenomenon actually obeys Benford's Law, and then to deduce (or suspect) unusual processes or departures from the norm by observing any possible changes from expected digital distribution, since such digital changes are always the features of some intrinsic deviation in the state of the phenomenon in question.

A list of 2175 common chemical compounds in use worldwide is given in the website http://www.convertunits.com/compounds/Z/. The selection of compounds was not made following any strict criteria, but rather simply by gathering information from many different relevant sources, and following the suggestions of users and chemists. The impressive variety in the list makes it appear to be a good and fair representative of any proper collection of chemical compounds in use for the purpose of Benford digital testing. It includes seemingly totally unrelated chemicals, such as those used in heavy industry, pharmaceutical, the food industry, and metallurgical plants, as well as many naturally occurring compounds. Certainly, no attention whatsoever was paid to molar

(molecular) mass in the process of selecting and compiling this list, at least not directly in any way. The molar mass is the combined or total mass of all the elements within the molecule. For example, the molar mass of the water molecule is 18.01528, having 2 hydrogen elements of 1.00794 weight each, and 1 oxygen element of 15.9994. Here are the 1st and 2nd digit distributions respectively of the molar mass in this list:

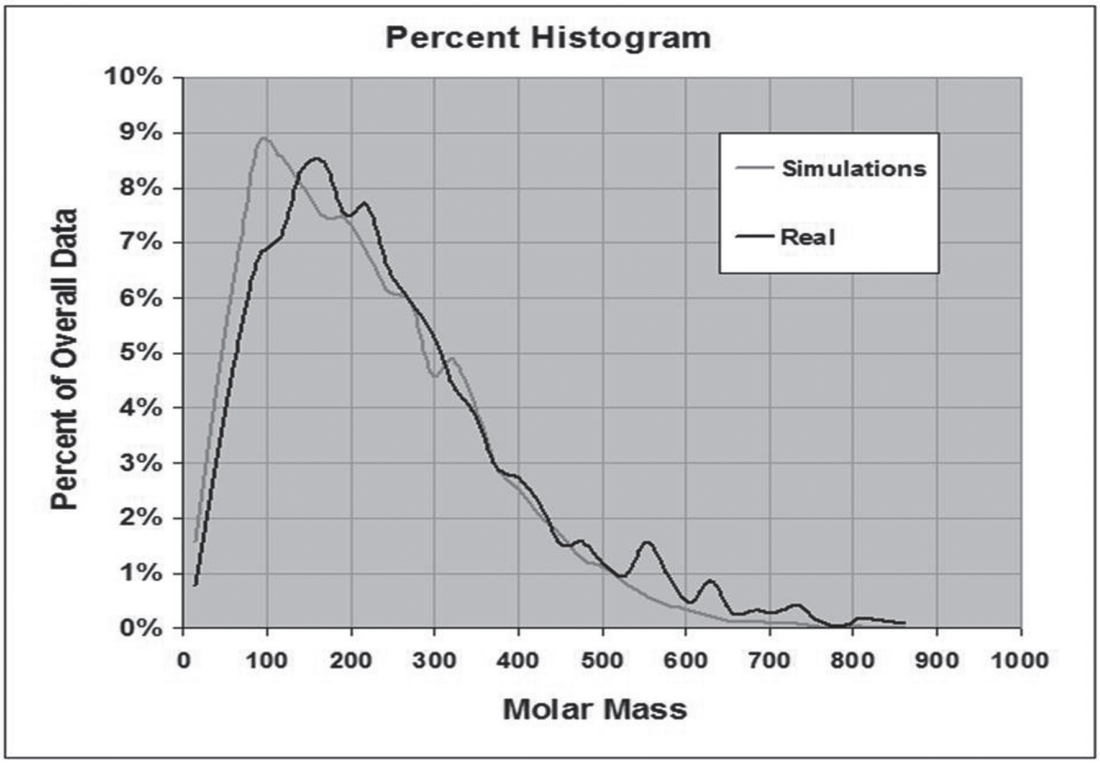
{31.9%, 25.2%, 16.1%, 8.4%, 5.7%, 4.3%, 2.9%, 3.2%, 2.3%}
{11.2%, 9.9%, 11.1%, 10.1%, 10.3%, 10.7%, 9.2%, 9.0%, 9.8%, 8.6%}

This is quite close to the logarithmic, and it constitutes one quite remarkable result! But why should molar mass be logarithmic? Curiosity compels one to compare this surprising result with a totally random selection, combination, and mixing from the Period Table. Would such blind and haphazard combination yield similar digital result? Monte Carlo Computer simulations of just such a scheme was performed, selecting randomly from the first 35 elements in the Period Table, ranging from the simple hydrogen, all the way to the element Bromine with its mass of 79.904 as the heaviest one in this simulation scheme. Bromine was arbitrarily chosen so as to avoid the heaviest of the elements which are rarely in use in chemistry. The first simulated element within this factitious or virtual molecule is then chosen as in the equal discrete uniform distribution {1 to 35}, and its frequency within the molecule is simulated separately from the discrete uniform {1 to 5}. The second element is chosen in likewise manner, independently, with possible duplication of elements and quantities. The third and final element is likewise simulated, but only once a coin is flipped showing a head (having 50% probability), otherwise, when a tail is found, it is aborted and the molecule is built just from the previous two selections. Thus the highest number of total elements a molecule can hope to have in such simulations is 15, while the least fortunate compound would have just 2 elements. The fact that most molecules in the real list of common compounds actually consist of 2 to 3 elements, and that each element usually appears no more than 5 times within a given compound, has provided the motivation for the above parameters and structure in the computer simulations. The 1st and 2nd digit distributions results of 10,000 such simulations came out very close to those of the list of real chemical compounds above, and are respectively:

{31.4%, 25.0%, 16.2%, 8.6%, 5.4%, 3.2%, 3.1%, 3.6%, 3.5%}
{11.7%, 11.1%, 11.2%, 10.7%, 10.2%, 9.4%, 9.9%, 8.7%, 8.6%, 8.5%}

Furthermore, the two adjusted histograms of the data sets themselves, of the type that puts both data sets on equal footing by listing percent of overall total (2175 & 10000) instead of simple counts, are remarkably similar, as shown in Figure 9. This strongly suggests that nature's way of combining elements together into molecules is done in almost totally random manner with regards to molar mass.

FIGURE 9
HISTOGRAM OF SIMULATED AND REAL DATA ON MOLAR MASS OF COMPOUNDS



A remarkable correspondence or analogy to the chemical molar mass data is found in company accounting data regarding revenues, namely the list of actual bills or receipts paid by customers. A typical purchase by a client shopping at an IT store might consists of 1 computer costing $860, 2 USB keys at $15 each, and 10 packages of CD disks at $5 a package. The total bill of $940 for all these products is simply derived from the random linear combination of the various prices listed at the shop, namely 1*($860) + 2*($15) + 10*($5). In this context, the individual mass of an element in the Periodic Table is analogous to the price of an item on sale in the list of prices, and the molar mass is analogous to the total bill paid the shopper. At play here are simultaneous random variables and decisions that determine the very limited selection of prices for products (typically only 1, 2, or 3) from that very long price list (typically in the hundreds) relating to all available products on sale, and with the possibility of buying multiple number of units of each product chosen. It is well-known that revenue data follows Benford's Law quite closely. Chemical molar mass and company's revenue data are only two particular manifestations of a much wider and general principle in Benford's Law called "Random Linear Combination" (RLC) which serves as yet another important cause for the prevalence of the logarithmic distribution in numerous real life data. Surely there are many other real life data sets having the same underlying statistical structure and therefore logarithmic as well, although this may not be immediately obvious when one contemplates such data types. For a convincing and remarkable demonstration of the strong effect Random Linear Combination has on digits, consider an extremely small shop selling only 6 items with its price list of {$1.25, $1.75, $6.75, $12.50, $35.00, $58.00}. Monte

Carlo computer simulations of bills (revenues) assuming hypothetical shoppers who buy exactly 2 items, while throwing two fair 6-sided dice to randomly decide on quantities bought, yield 1st digits of: {30.8%, 19.6%, 14.0%, 9.7%, 5.5%, 6.6%, 6.2%, 5.1%, 2.6%}!

Three essential statistical distributions widely used in quantum mechanics and thermodynamics, Boltzmann-Gibbs, Fermi-Dirac, and Bose-Einstein, have been found by Lijing Shao et al. to be nearly (and at times exactly) logarithmic, which implies that Benford's Law is quite prevalent in those fields. Moreover, they conjectured that Benford's Law itself might be (or might hint at) a truly profound and fundamental law in nature in general, and not only in physical statistics, perhaps representing a more fundamental principle behind the complexity of nature, governing the properties of many physical systems.

IX. THE CASE OF K/X DISTRIBUTION

The probability density function f(x) = k/x occupies a central position and importance in the study of Leading Digits and Benford's Law, and studying all aspects of this distribution is essential for a complete understanding of the phenomena.

Consider the probability density function of the form f(x)=k/x over the interval $[10^S,10^{S+G}]$ where S is any real number, G is any positive integer, and k is a constant depending on the values of S and G. It shall be shown that the sum of all the areas under the curve where digit d leads is indeed $LOG_{10}(1+1/d)$, namely perfectly logarithmic. Let us prove this assertion. We first note that the entire area should sum to one, that is

$\phi k/x\, dx = 1$ over $[10^S,10^{S+G}]$, therefore $k[\ln(10^{S+G}) - \ln(10^S)] = 1$, or

$k[(S+G)\ln10 - (S)\ln10] = 1$, so that $k*\ln10*[(S+G)-(S)]=1$ and

$k* \ln10*G = 1$, so $k = 1/[G*\ln10]$. Notice, that this determination of k was in total generality, where G can assume any value, not necessarily only an integer, and that G represents the difference in the base 10 exponents of the two boundaries spanning the entire interval length of x in question.

Secondly, given a particular p.d.f. of the form k/x on (a, b), we note that for any two subintervals of (a, b) having the same exponent difference, their areas under the curve are identical. Given $[10^P, 10^{P+I}]$ and $[10^Q, 10^{Q+I}]$, both contained inside (a, b), P and Q being <u>any</u> set of numbers, not necessarily integers, we know that the values of their related constant k are identical since they belong to the same distribution defined on

(a, b). The areas under the curve are $k[\ln(10^{P+I}) - \ln(10^P)]$ and $k[\ln(10^{Q+I}) - \ln(10^Q)]$ respectively, or simply $k[(P+I)\ln10 – (P)\ln10]$ and $k[(Q+I)\ln10 – (Q)\ln10]$, simplifying we get: $k*\ln(10)*[(P+I) – (P)]$ and $k*\ln(10)*[(Q+I) – (Q)]$, which yields $k*\ln10*I$ as the same area for each subinterval. If the whole interval (a, b) is expressed as $[10^S,10^{S+G}]$ so that $k = 1/[G*\ln10]$ then area for each is simply I/G, namely the ratios of exponent difference of the subinterval to the exponent difference of the entire range.

For example, for k/x defined on (1, 10000), [1, 10] and [100, 1000] have equal areas,

[1, 10] is narrower on the x-axis but with a very high p.d.f. value, while [100, 1000] is extremely long on the x-axis in comparison, but it comes with a much shorter p.d.f. value. This trade-off cancels out each effect exactly so that areas end up the same.

To set the stage for the proof, we represent the interval in question here as

$[10^{N+f},10^{N+f+G}]$, where N is zero or some positive integer, f is some possible fractional part, and G is a positive integer representing the integral part of the difference in exponents.

We initially let f = 0, a restriction that would be relaxed later. Considering any digit D, the probability that D leads is given by the area under f(x)=k/x on the following intervals:

{[$D10^N$, $(D+1)10^N$], [$D10^{N+1}$, $(D+1)10^{N+1}$], ...**G times**…[$D*10^{(N+G-1)}$,$(D+1)*10^{(N+G-1)}$]}.

This is so because it is on these intervals and these intervals alone that D leads.

Calculating the various definite integrals we obtain:

$k[\ln((D+1)10^N) - \ln(D10^N)] + k[\ln((D+1)10^{N+1}) - \ln(D10^{N+1})]$

$+….$**G times**$….+ k[\ln((D+1)10^{N+G-1}) - \ln(D10^{N+G-1})]$ or:

$k[\ln(D+1)+N*\ln(10)-\ln(D)-N*\ln(10)\ ] + k[\ln(D+1)+(N+1)*\ln(10)-\ln(D)-(N+1)*\ln(10)]$

$+ ….$**G times**$…. + k[\ln(D+1)+(N+G-1)*\ln(10)-\ln(D)-(N+G-1)*\ln(10)]$

Canceling out like terms (not involving D) we are left with:

$k[\ln(D+1)-\ln(D)] + k[\ln(D+1)-\ln(D)]+$ **…(G times) ... +** $k[\ln(D+1)-\ln(D)]$

$k[\ln[(D+1)/(D)]] + k[\ln[(D+1)/(D)]]+$ **…(G times) ... +** $k[\ln[(D+1)/(D)]]$

That is: $G*k*\ln[(D+1)/(D)]$. Substituting here the expression for k above we obtain:

$G*(1/[G*\ln10])*\ln[(D+1)/(D)] = \ln[(D+1)/D]/\ln10$. This expression uses the natural logarithm base e, and applying the logarithmic identity $LOG_AX=LOG_BX / LOG_BA$

to convert this ratio to the common base 10 we get:

$LOG_{10}[(D+1)/D]/LOG_{10}[e] / LOG_{10}[10]/ LOG_{10}[e]$

$LOG_{10}[(D+1)/D]/LOG_{10}[e] / 1/ LOG_{10}[e]$

$LOG_{10}[(D+1)/D]$

$LOG_{10}[1+1/D]$

Let us now see why $f\neq0$ should not yield any different result. As shown above k depends solely on the difference between the exponents, being that its expression is

$k = 1/[G*\ln10]$, hence since $f\neq0$ (a slight shift to the right from the f=0 situation) does not alter that difference here, and k is still of the same value (whether f equals to zero or not), and thus function x/k is not altered either. Now, since areas under the curve are identical for any two subintervals of the same exponent difference, it follows that any nonzero f that requires an additional area to the right of the intervals

{[$D10^N$, $(D+1)10^N$], [$D10^{N+1}$, $(D+1)10^{N+1}$], … , [$D*10^{(N+G-1)}$,$(D+1)*10^{(N+G-1)}$]}.

would then also require an identical subtraction on the left side of that intervals! This completes the proof.

X. DETECTING ACCOUNTING FRAUD

In the early 1990s it was first suggested to apply the digital property of Benford's Law as a technique in forensic data analysis of accounting and financial data to detect fraud.

Following this innovation soon afterwards, it has been increasingly used by accounting firms, governmental tax authorities such as the IRS in the USA, and now in most other tax authorities worldwide, as routine check on data. It is important to recognize the fact that each well-defined piece of data has its own particular leading digits signature, a sort of a hidden digital code - not immediately obvious during the first visual (preliminary) inspection of it when the focus is on numbers and quantities but not on their digital language. Fraud, anomalies, data entry errors, and irregularities can be detected using Benford's Law by comparing the actual distribution of the first digits in a set of accounting or financial data to the theoretical distribution given by Benford's Law. Maliciously invented fake data obviously does not obey Benford's Law, but instead digits appear all equally likely just as most people would mistakenly intuit. A cautionary flag is raised if deviation of actual from theoretical is significant, which calls for further scrutiny and examination of data.

As an example we look at some hypothetical accounting data from 5 different Costa Rican companies (firms) representing revenues, that is, sales receipts:

FIGURE 10
ACCOUNTING DATA FOR FIVE COSTA RICAN COMPANIES

Firm A	Firm B	Firm C	Firm D	Firm E
624.00	120.30	182.03	420.40	13.00
149.00	74.40	158.11	74.00	62.69
104.74	107.71	37.62	3.54	19.50
171.00	17.43	363.60	645.00	221.05
102.26	9.99	361.99	211.40	27.05
179.98	373.68	150.00	8.40	36.75
9.77	209.00	209.87	143.24	11.25
373.87	14.09	250.61	23.87	56.94
23.48	54.00	3.62	5.64	225.00
12.98	219.00	575.44	503.24	17.55
480.61	3.62	79.45	978.20	10.75
4.37	636.20	245.53	10.87	14.75
116.25	1332.81	404.84	43.80	493.05
149.00	174.38	86.77	3.84	120.84
274.92	32.40	114.35	8.97	33.45
89.00	225.00	119.89	935.70	11.42
20.70	28.75	19.10	7.25	100.00
224.93	479.00	6.62	54.30	80.00
50.00	3.62	165.75	29.35	20.75
662.75	529.75	25.10	74.60	50.00

Initially, by just looking at the numbers, we can not detect anything unusual or wrong. Yet, forensically investigating 1st leading digit distributions, we get this result:

FIGURE 11
LEADING DIGITS OF THE ACCOUNTING DATA OF FIVE CR COMPANIES

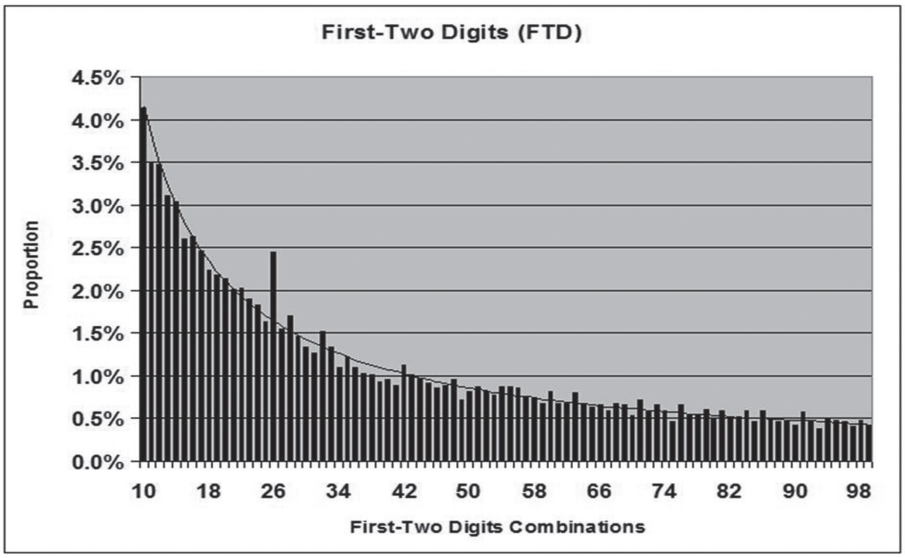
Digit	Firm A	Firm B	Firm C	Firm D	Firm E
1	40%	30%	35%	10%	45%
2	20%	20%	20%	15%	20%
3	5%	20%	20%	10%	10%
4	10%	5%	5%	10%	5%
5	5%	10%	5%	15%	10%
6	10%	5%	5%	5%	5%
7	0%	5%	5%	15%	0%
8	5%	0%	5%	10%	5%
9	5%	5%	0%	10%	0%

Now when data is visualized digitally it is obvious that Firm D comes under some strong suspicion since it gives approximately uniform digital distribution. Data of all other companies are not exactly Benford but show a clear and decisive approximation of it. For better forensic result it is recommended to include 4 digital tests:

1) First digits distribution.
2) Second digits distribution.
3) Combination of the first-two digits distribution.
4) Combination of the last-two digits distribution.

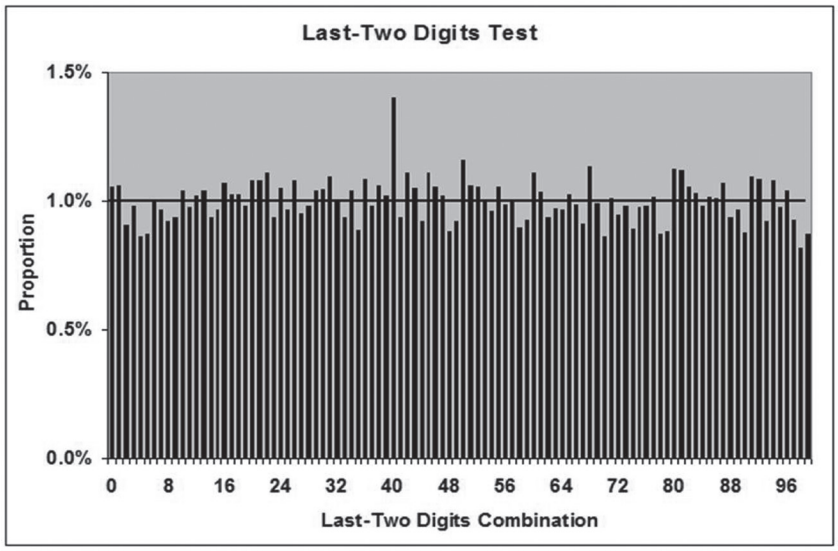
The following is a chart of the First-Two Digit test performed on a hypothetical company:

FIGURE 12
FIRST-TWO DIGIT TEST PERFORMED ON A HYPOTHETICAL COMPANY



The continuous line is the theoretical Benford proportion, LOG $_{10}$ (1 + 1/pq), which falls off gradually from 4.1% to 0.4%. There is a material issue for the auditors to examine further here because of the substantial spike at 26. It means that there were excess amounts starting with 26, such as $26,800, $260, or $260,598. This needs to be audited and examined in more details. Since the issue here with the 26 spike would not have been detected employing just the 1st digit distribution, it is recommended always to look also into FTD. The following is a chart of the Last-Two Digit test performed on a hypothetical company:

FIGURE 13
LAST-TWO DIGIT TEST PERFORMED ON HYPOTHETICAL COMPANY



The continuous line is the theoretical Benford proportion, which is steady at 1/100, because it gives equal proportion to all 100 possibilities of {00, 01, 03, … , 97, 98, 99}. There is a material issue

for the auditors to examine further in this accounting data set because of the substantial spike at 40. It means that there were excess amounts ending with 40, such as $19,755.40, $81.40, or $46.40. This needs to be audited and examined in more details.

What part of the data do we examine? We never examine sums, aggregates, summaries, and totals in digital analysis. Rather we look into actual (original/raw) expenses, revenues, account receivables, and so forth, because they follow Benford's Law, not totals!

Unlike typical statistical methods and practices regarding data itself where samples are more economical and practical to take, or are the only possible choice in surveys and studies, here for digital examinations it is preferred to examine ALL the data available relating to the accounting test in the audit. This is so because there is more accuracy and reliability for large data sets than smaller ones in digital analysis. For statistical significance, the test is performed only if the company has a lot of values, therefore the more values the better. A very small company with few entries would not be suitable for digital analysis using Benford's Law, because Type I errors would occur too frequently. Type I error, also known as "false positive", occurs when digital analysis mistakenly concludes that an honest company is fraudulent. The standard practice is to eliminate from the data set all negative values, all zero values, as well as all low values less than 10, before any digital analysis is done. There are two reasons for eliminating values less than 10; (i) they are not so important to the auditor, (ii) their 2nd digit is often 0 since $7 is written as $7.00 therefore artificially increasing digit 0.

It is important to note that when a company is suspected of fraud due to a failure to pass Benford's digital test, we still do NOT know which entries were fake, and which were honest. Perhaps all were fake. All we know is that the tax report in its entirety is likely to be dishonest. If a company has tens of thousands of entries in its tax report, and gave only a few fraudulent values, say just 23, it would be impossible to discover such fraud in digital analysis. The challenge arises when these 23 fake values are of very high values representing a large proportion of money involved, yet not being detected at all, in short: having false negatives. Fortunately, almost all cheaters and fraudsters do not know about the existence of Benford's Law (at this present epoch at least) and so they may prefer to cheat by changing thousands of numbers instead of only 23, mistakenly thinking that by spreading the fraud around many numbers this will not be detected.

When a company wants to report strong income to attract investors for example, then amounts such as 797,156 and 29.9 are changed and rounded up and reported as 800,000 and 30. This would artificially increase the proportion of the 0 digit in any digital analysis. On the other hand, when a company wants to under-report profits in order to pay less tax it would round down profits, and this will result in artificially reducing the 0 digit in the analysis. Hence over-representation of digit 0 indicates an attempt to increase amounts, while under-representation of it indicates an attempt to decrease amounts. In general, an excess in the 0 digit proportion indicates that some fake rounding of amounts occurs.

For super large account such as ones with more than 100,000 values for example, forensic digital analysis might calls for the examination of too many numbers (amounts) thought to be suspicious, which is too costly and expensive for the company. In this case is it better to apply the First-Three-Digit test utilizing the theoretical probability LOG_{10} (1 + 1/pqr) followed by the Value Duplication test discussed below, because then the focus of analysis is narrowed down to fewer numbers (amounts), thus lowering the expense of the examination.

Clearly, any fraud due to unreported dealings where no transaction is actually recorded in company data, such as bribes, kickbacks, asset thefts, and so forth, can not be detected by digital analysis. In addition, very few (rare) accounting data types do not obey Benford's Law to begin with and therefore can not be so tested, such as payroll (salaries) amounts, amounts with an arbitrary minimum or maximum, or amounts that are influenced by human thought like ATM withdrawals and fixed prices.

Post statistical result:

If the company's data is found not to obey Benford's Law, then there are 4 possibilities:

1. False positive. The company is honest, but by some rare random statistical chance its data deviated from Benford.
2. The assumption that this particular type of accounting data follows (or should follow) Benford's Law is not really true, and the company is actually honest.
3 The company has some particular items under sale that causes its revenues to favor certain digits. For example, if its main product is a very popular laptop with a price of $799, then digit 7 would be over-represented, and digital tests would always indicate possible fraud, but the company is actually honest.
4. The company is dishonest, and the data is fraudulent – faked by its accountants.

XI. STATISTICAL TESTS IN DIGITAL DATA ANALYSIS

Both, the Z Test and the chi-sqr Test in this section are applicable only in those ideal situations where the statistician is drawing a truly random sample from some population data to determine whether the population is logarithmic or not. In reality, the accountant or the auditor performing forensic digital analysis in fraud detection environment can not assume that, because the data itself here is indeed the whole population in question. Unfortunately, both tests have been used, and are still being used erroneously to test for Benford compliance. There are two types of tests: (I) An overall test taking all the digits into account by combining all 9 (or more) deviations from expected proportions. (II) A digit-by-digit test (separately for each digit), where perhaps some appear deviant and suspicious and some appear correct as expected. The latter test gives more specific information about which digits are off, but Type I error (false positive) is about 7 times more likely than for the overall test. The relevant digits in question could be: 1 to 9 for the 1st order; 0 to 9 for the 2nd order; 10 to 99 for the first-two digits order, and 00 to 99 for the last-two digits test.

Note that for digital forensic analysis in the context of Benford's Law, the significance level of 5% is traditionally used. The strict level of 1% is not considered appropriate here.

Z Test:

H_0: Data obeys Benford's Law in the context of the particular digit i.
Pe = the expected Benford's proportion log(1+1/i) for the particular digit i in question.
Po = the observed (actual) proportion of numbers being led by digit i in the data set.
N = the number of observations (money values) in the data set of the particular account.
The standard deviation SDi for each particular i digit's expected proportion is:
SDi = Square Root [Pe(1-Pe) / N] for digit i in (1-9) or (0-9) or (10-99)
This Z test is performed on one particular digit i (individually) with:
Zi statistic = (|Po – Pe| - 1/(2N)) /SDi

Reject the Null Hypothesis H_0 at the p% confidence level if Zi value is larger than Z with p% critical value. Z refers to the Standardized Normal Distribution, that is N(0, 1).

The term 1/(2N) is the continuity correction factor and is used only when it is smaller than the absolute value term (hence Zi is never a negative quantity).

An alternative expression for the Z statistic is:
Zi statistic =(√N)* (|Po – Pe| - 1/(2N)) /(√(Pe (1 - Pe)))

A major difficulty with the Z-test in the context of Benford's Law is that when N is large the test becomes too sensitive and even very small deviation from the Benford proportion are flagged as significant (false positive). In short: the z-test suffers from excess power. Generally speaking, any account with over 100,000 entries is considered too large for the Z-test as well as for the chi-sqr test.

Chi-sqr Test:

H_0: The data set obeys Benford's Law overall considering the set of all relevant digits in question, namely for digits (1-9), or (0-9), or (10, 99).

Chi-sqr statistic = $N^*\sum [(Pe - Po)^2/Pe]$ summed over digits (1-9) or (0-9) or (10-99)

Reject the Null Hypothesis H_0 at the p% confidence level if chi-sqr value is larger than chi-sqr-p% critical value with (RD-1) degrees of freedom. RD is the number of Relevant Digits in the particular test. For example, RD=9 for the 1st digits, RD=10 for the 2nd digits, RD=90 for the first-two digits, and RD=100 for the last-two digits.

Here again the major difficulty is when N is large (>100,000 approximately) and the test becomes too sensitive that even very small deviations from the Benford proportions flags the entire data set as significantly non-logarithmic (false positive). In short: chi-sqr test suffers from excess power, and its use by auditors is almost always erroneous!

Note that rejecting the null hypothesis via the chi-sqr test does not tell us specifically which digits are problematic and which are not, which are over-represented and which are under-represented.

Value Duplication Test:

This is a test to identify the specific numbers that were causing the spikes on the 1st digits, 2nd digits, first-two digits, and last-two digits graphs. It frequently helps to discover some specific numbers occurring abnormally too often. To perform this test a table is created (typically in MS-Access or other Database software) showing all values and their frequencies of occurrence. This table is typically ordered high to low for easy inspection.

XII. CONCLUSIONS

Examination of digital distributions in data provides a practical new technique in forensic data analysis with regards to authenticity or falseness of data. The use of these digital forensic tests have been spreading rapidly in the past 15 years and have become the standard procedures in most Tax Revenue Departments of governments worldwide, as well as in large accounting and auditing companies. There are some limits and difficulties applying them at times, especially when very few invented values of large and significant amounts are inserted into a large authentic data set, resulting in low concentration of false values. In addition, a minority of data types does not obey Benford's Law in the first place, such as salary accounts for example, and therefore these primary digital tests are not available for forensic analysis there. Yet, apart from these two examples and a few other cases, the vast majority of everyday real data obeys Benford's Law, or at least is very close to it, a fact which enables the statistician to apply the law in almost all financial and accounting situations. Another interesting application in this context is the possibility of forensic data analysis of official election results, testing for the possibility of having a fraudulent democracy. Since population data by city or by province are almost perfectly Benford, so should be electoral data result, which is simply its breakdown by the fractions of the relevant political parties.

BIBLIOGRAPHY

Benford Frank (1938), "The law of anomalous numbers," Proceedings of the American Philosophical Society, 78 p. 551.

Burton David (1993) "The history of mathematics: an introduction". Third Edition. On the Golden Ratio in Pages 57, 117, 166, 267, and 268.

Cho Wendy, et al. (2007) "Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance", Wendy Cho, Joanne Lee, George Judge, The American Statistician.

Cleary Richard, Thibodeau Jay (2005) "Applying Digital Analysis Using Benford's Law to Detect Fraud: The Dangers of Type I Errors", Auditing: A Journal of Practice & Theory.

Durtschi Cindy, et al. (2004) "The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data", Journal of Forensic Accounting, 1524-5586/Vol V. 2004, pp 17-34

Flehinger B.J. (1966) "On the Probability that a Random Integer has Initial Digit A," American Mathematical Monthly, Vol 73, No.10, Dec., 1966, 1056-1061

Hill Theodore (1995) "A statistical derivation of the significant-digit law," Statistical Science 10(4), 1995c.

Hill Theodore (1998) "The First Digit Phenomena," American Scientist, July-August 1998, v86, n4, p358(6).

Kossovsky Alex Ely (2006) "Towards a Better Understanding of the Leading Digits Phenomena (Benford's Law)", Cornell University Library, http://arxiv.org/abs/math/0612627.

Lijing Shao, Bo-Qiang Ma (2010) "The Significant Digit Law in Statistical Physics", http://arxiv.org/pdf/1005.0660v1.pdf, Peking University, China, 5 May 2010

Miller Steven (2008) "Chains of distributions, hierarchical Bayesian models and Benford's Law", Jun 2008, http://arxiv.org/abs/0805.4226.

Newcomb Simon (1881) "'Note on the Frequency of Use of the Different Digits in Natural Numbers," American Journal of Mathematics, 4 (1881): 39-40.

Pinkham Roger (1961) "On the Distribution of First Significant Digits," The Annals of Mathematical Statistics, 1961, Vol.32, No. 4 , 1223-1230.

Raimi Ralph A. (1969) "The Peculiar Distribution of First Digit", Scientific America, Sep 1969.

Raimi Ralph A. (1976) "The First Digit Problem", American Mathematical Monthly, Aug-Sep 1976.

Raimi Ralph A. (1985) "The First Digit Phenomena Again," Proceedings of the American Philosophical Society, Vol. 129, No 2, June, 1985, 211-219.

Sambridge Malcolm, et al. (2010) "Benford's Law in the Natural Sciences", M. Sambridge, H. Tkalcic, A. Jackson, Geophysical Research Letters, Vol. 37, L22301

Sambridge Malcolm, et al. (2011) "Benford's Law of First Digits: From Mathematical Curiosity to Change Detector", Sambridge Malcolm, Hrvoje T. and Arroucau Pierre, Asia Pacific Mathematics Newsletter