Properties of Data Sets that Conform to Benford's Law

Author: R.C. Hall: MSEE, BSEE
e-mail: rhall20448@aol.com

Abstract

The concept and application of Benford's law have been examined a lot in the last 10 years or so, especially with regard to accounting forensics. There have been many papers written as to why Benford's law is so prevalent and the concomitant reasons why (proofs). There are, unfortunately, many misconceptions such as the newly coined phrase "the Summation theorem", which states if a data set conforms to Benford's law then the sum of all numbers that begin with a particular digit (1,2,3,4,5,6,7,8,9) should be equal. Such is usually not the case. For exponential functions ($y = a^{x}$) it is but not for most other functions. I will show as to why this is the case. The distribution tends to be a Benford instead of a Uniform distribution.

Also, I will show that if the probability density function (pdf) of the logarithm of a data set begins and ends on the x axis and if the values of the pdf between all integral powers of ten can be approximated with a straight line then the data set will tend to conform to Benford's law.

What is Benford's Law

Benford's law is a product of a number theory concept relating to the distribution of numbers, more specifically, the lead digits (or first two digits) of numbers obtained from a data set. For example, for a group of numbers such as: 23178, 56789, 32150, and 09876 the lead digits are, respectively, 2, 5, 3, and 9 (leading zeros are excluded) the first two leading digits would be 23, 56, 32, and 98. One would expect the distribution of the aforementioned leading digits (1,2,3,4,5,6,7,8,9 or 10, 11, 12 ……97,98,99) to be a Uniform distribution (equiprobable) but such is usually not the case. Most numbers that occur in nature consist of more of the lower digits than the higher digits as first digits.

About 30% of all first digits begin with the number 1 and almost half of all of these numbers begin with either 1 or 2. One would expect the number 1 to appear (1/9 or 11.11%) of the time and either 1 or 2 to appear about (2/9 or 22.22%) of the time. Only about 4.5% of these numbers start with 9 instead of the expected 11.11%. More specifically the percentages are as follows:

First Digit	Percentage of Occurrence
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%

More exactly:

First Digit	Fractional value of Occurrence
1	LOG10(2)
2	LOG10(3/2)
3	LOG10(4/3)
4	LOG10(5/4)
5	LOG10(6/5)

6	LOG10(7/6)
7	LOG10(8/7)
8	LOG10(9/8)
9	LOG10(10/9)

Pr digit(n) = LOG10 ($\frac{n+1}{n}$) ; n = 1,2,3,4,5,6,7,8,9

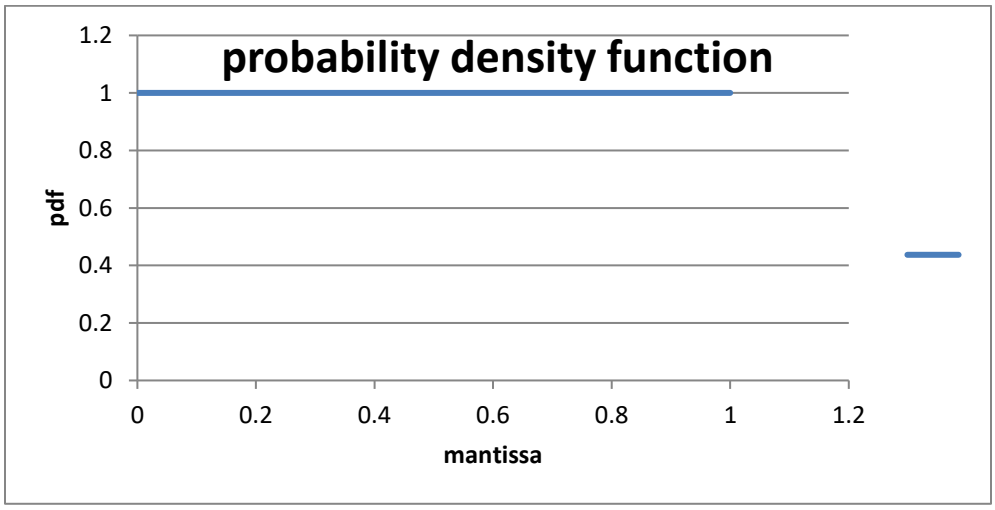
Examples of data that follow Benford's law are population of cities, molecular weights, accounting data, tax return data, and multiplication of random numbers.

Numbers such as assigned numbers i.e. check numbers, telephone numbers, invoice numbers follow a Uniform distribution instead (one sequential number per entity only).

Why is Benford's Law true?

Benford's law (probability of 1st digits (n) = LOG10($\frac{n+1}{n}$)) is predicated on a Uniform distribution of the mantissas of the logarithms of a data set**. If the probability density function of the mantissas of the logarithm of a data set is a Uniform distribution then the data set conforms exactly to Benford's law.**

Proof:



Fig#1 – probability density function of the logarithm of a data set

Given: pdf(log x) = 1.0

1) y = log(x)
2) pdf(y) dy = pdf(x) dx
3) pdf(x) = pdf(y)$\frac{dy}{dx}$
4) y = log(x) = $\frac{\ln(x)}{\ln(10)}$
5) $\frac{dy}{dx}$ = $\frac{1}{\ln(10)x}$
6) Pdf(x) = $\frac{pdf(y)}{\ln(10)x}$ = $\frac{1}{\ln(10)x}$

Probability distribution function = $\int_1^x pdf(x)dx$

$\int_1^2 pdf(x)dx = \int_1^2 \frac{dx}{\ln(10)} dx = \frac{\ln(2)}{\ln(10)}$ = Log(2)

$\int_2^3 pdf(x)dx = \int_2^3 \frac{dx}{\ln(10)} dx = \frac{\ln(3/2)}{\ln(10)}$ = Log(3/2)

$\int_3^4 pdf(x)dx = \int_3^4 \frac{dx}{\ln(10)} dx = \frac{\ln(4/3)}{\ln(10)}$ = Log(4/3)

$\int_4^5 pdf(x)dx = \int_4^5 \frac{dx}{\ln(10)} dx = \frac{\ln(5/4)}{\ln(10)}$ = Log(5/4)

$\int_5^6 pdf(x)dx = \int_5^6 \frac{dx}{\ln(10)} dx = \frac{\ln(6/5)}{\ln(10)}$ = Log(6/5)

$\int_6^7 pdf(x)dx = \int_6^7 \frac{dx}{\ln(10)} dx = \frac{\ln(7/6)}{\ln(10)}$ = Log(7/6)

$\int_7^8 pdf(x)dx = \int_7^8 \frac{dx}{\ln(10)} dx = \frac{\ln(8/7)}{\ln(10)}$ = Log(8/7)

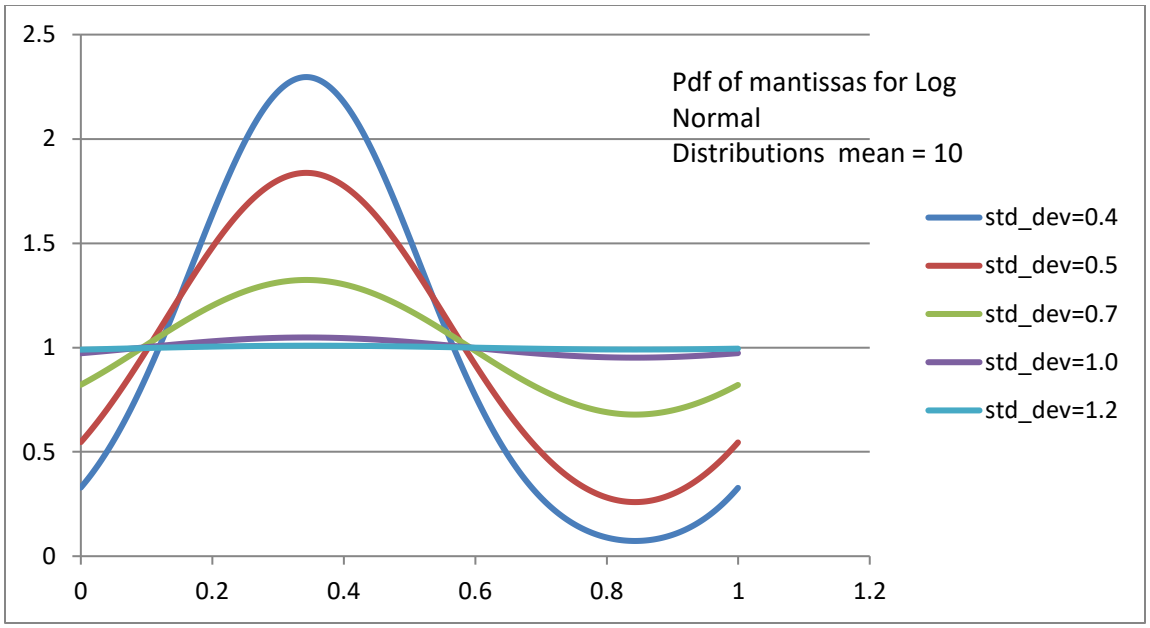
$\int_8^9 pdf(x)dx = \int_8^9 \frac{dx}{\ln(10)} dx = \frac{\ln(9/8)}{\ln(10)}$ = Log(9/8)

$\int_9^{10} pdf(x)dx = \int_9^{10} \frac{dx}{\ln(10)} dx = \frac{\ln(10/9)}{\ln(10)}$ = Log(10/9)

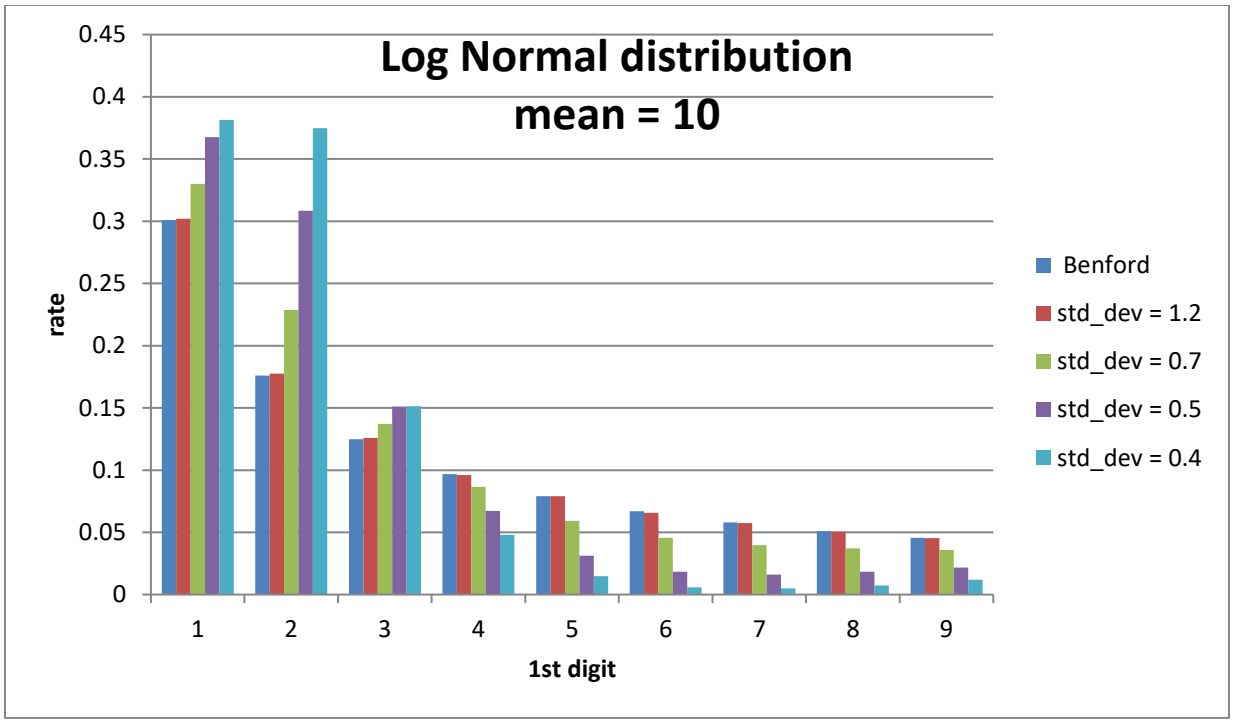
It can be further postulated that there is a rough proportionality that exists between the degree of uniformity of the probability density function (pdf) of the mantissas of the logarithm of a data set and compliance with Benford's law

It will be proven and demonstrated later that a Log Normal distribution approaches a Benford distribution as its standard deviation approaches infinity and approaches a Normal distribution as its standard deviation approaches zero.

For each degree of mantissa pdf uniformity I have plotted, with the aid of Microsoft Excel, the corresponding 1st digit distribution. The results clearly demonstrate that as the degree of mantissa pdf uniformity decreases the less the data conforms to Benford's law. With a mean of 10 and a standard deviation of 2.0 there is almost complete conformity whereas with a mean of 10 and a standard deviation of 0.2 there is virtually no conformity



Fig#2- Sum of the pdf of each mantissa value for each corresponding IPOT for a Log Normal distribution with a mean of 10 and various standard deviations



Fig#3- 1st digit distribution for a Log Normal distribution mean = 10 for various standard deviations

Scale invariance

The scale invariance associated with Benford's law states that if the original data were multiplied by a constant the Ist digits distribution would stay the same. An example of this would be converting data from inches to centimeters. The following argument constitutes a proof of this assertion.

Let a = scale factor: $\frac{1}{\ln(10)} \int_{a10^N}^{a10^{N+1}} \frac{dx}{x} = \frac{1}{\ln(10)} [\ln(a10^{N+1}) - \ln(a10^N)] =$

$\frac{1}{\ln(10)}[(N+1)\ln(10) + \ln(a) - N\ln(10) - \ln(a)] = \frac{1}{\ln(10)}[N\ln(10) - N\ln(10) + \ln(a) - \ln(a) + \ln(10)] = \frac{\ln(10)}{\ln(10)} = 1$

Numbers starting from a →2a: $\frac{1}{\ln(10)} [\int_{a10^N}^{2a10^N} \frac{dx}{x}] = \frac{\ln(2a10^N) - \ln(a10^N)}{\ln(10)} =$

$\frac{\ln(2)+\ln(a)-\ln(a)+n\ln(10)-n\ln(10)}{\ln(10)} = \frac{\ln(2)}{\ln(10)} = \log_{10} 2$

Likewise for numbers starting with 2: $\frac{1}{\ln(10)}[\int_{2a10^N}^{3a10^N} \frac{dx}{x} = \frac{\ln(\frac{3}{2})}{\ln(10)} = \log_{10} 3/2$

Example: converting inches to centimeters

Scale factor: a = 2.54 centimeters/inch

$\frac{1}{\ln(10)} [\int_{2.54\,x10^N}^{2x2.54x10^N} \frac{dx}{x} = \frac{1}{\ln(10)}[\ln(2.54) + \ln(2) + N\ln(10) - \ln(2.54) - N\ln(10)] = \frac{\ln(2)}{\ln(10)} = \log_{10} 2$

m = 1…….9 $\frac{1}{\ln(10)} \int_{am10^N}^{a(m+1)10^N} \frac{dx}{x} = [\ln(a(m+1)x10^N) - \ln(amx10^N)]/\ln(10) =$

$\ln(a) + \ln(m+1) + N\ln(10) - \ln(a) - \ln(m) - N\ln(10) = [\ln(m+1) - \ln(m)]/\ln(10) = \ln(\frac{m+1}{m})/\ln(10) = \log_{10} \frac{m+1}{m}$

Exponential functions

Exponential functions such as $y = a^r$ conform exactly to Benford's law. It can easily be proven that the probability density function of an exponential is $\frac{1}{\ln(10)x}$ and, therefore, completely conforms to Benford's law.

Proof that an exponential function conforms to Benford's Law.

1) Let exponential function $y = 10^x$

2) Let $v = \text{Log}_{10}(y) = x\,\text{Log}_{10}(10) = x$, which is the probability distribution function of the log of 10^x as the log of 10^x varies from 0 to 1

3) The probability density function of the log of 10^x is the derivative of v with respect to x, which is 1.

4) Apply the formula $pdf_v\, dv = pdf_y\, dy$

5) $pdf_y = pdf_v \times \frac{dv}{dy}$

6) $v = \text{Log}_{10}(y) = \frac{\ln(y)}{\ln(10)}$

7) $\frac{dv}{dy} = \frac{1}{y\ln(10)}$

8) $pdf_y = \frac{1}{y\ln(10)}$

9) $\int_a^b pdf_y\, dy$ = Probability [Pr($a \le y \le b$)] =

10) $\int_a^b \frac{dy}{y\text{Ln}(10)} = \frac{1}{\text{Ln}(10)} \int_a^b \frac{dy}{y} = \frac{\text{Ln}(b) - \text{Ln}(a)}{\text{Ln}(10)} = \frac{\text{Ln}\frac{b}{a}}{\text{Ln}(10)} = \text{Log}_{10}(\frac{b}{a})$

11) Let b = 2, a = 1; $\text{Log}_{10}(2) = 0.30103$

Let b = 3, a = 2; $\text{Log}_{10}(\frac{3}{2}) = 0.176091$

Let b= 4, a = 3; $\text{Log}_{10}(\frac{4}{3}) = 0.124939$

Let b = 5, a = 4; $\text{Log}_{10}(\frac{5}{4}) = 0.096910$

Let b = 6, a = 5; $\text{Log}_{10}(\frac{6}{5}) = 0.079181$

Let b = 7, a = 6; $\text{Log}_{10}(\frac{7}{6}) = 0.066947$

Let b = 8, a = 7; $\text{Log}_{10}(\frac{8}{7}) = 0.057992$

Let b = 9, a = 8; $\text{Log}_{10}(\frac{9}{8}) = 0.051153$

Let b =10, a = 9; $\text{Log}_{10}(\frac{10}{9}) = 0.045757$

The Ist digit distribution conforms to Benford's Law.

Log Normal Distributions:

As already having been illustrated with the Log Normal probability density function as the standard deviation approaches infinity the probability density function (pdf) of the mantissas of the logarithm of a data set approaches a uniform distribution and, therefore, approaches a Benford distribution. The following explanation is proof of this assertion.

Proof that as the standard deviation of a Log Normal distribution approaches infinity the distribution becomes a Benford distribution i.e. the probability density function approaches k/x

1) The Benford probability density function = 1/xln(10).
2) The Log Normal probability density function = $\frac{1}{x\sqrt{2\pi\sigma^2}}$ $e^{-(\mathrm{Ln}(x)-u)^2/2\sigma^2}$
3) For x=1: 1/xln(10) = 1/ln(10); $\frac{1}{x\sqrt{2\pi\sigma^2}}\, e^{-(\mathrm{Ln}(x)-u)^2/2\sigma^2} = \frac{1}{\sqrt{2\pi\sigma^2}}$ $e^{-(u)^2/2\sigma^2}$
4) Normalize by multiplying the Log Normal distribution by $\frac{\sqrt{2\pi\sigma^2}}{\ln(10)}\, e^{(u)^2/2\sigma^2}$

5) The difference between the two distributions is : $\frac{1}{x\ln(10)}$ - $\frac{1}{x\ln(10)}$ ($e^{-(\mathrm{Ln}(x)-u)^2/2\sigma^2}$) =

6) $\frac{1}{x\ln(10)}$ (1-$e^{-(\mathrm{Ln}(x)-u)^2/2\sigma^2}$)
7) For any given value of x the value 1-$e^{-(\mathrm{Ln}(x)-u)^2/2\sigma^2}$ approaches 0; since $e^{\frac{k(\mathrm{constant})}{\sigma^2}}$ approaches 1 as σ approaches ∞.

Also, as the standard deviation of a Log Normal distribution approaches 0, the Log Normal probability density function approaches a Normal or Gaussian distribution with a mean of e^u and a standard deviation of σ e^u where u is the mean of the Log Normal pdf and σ is the standard deviation of the same Log Normal pdf. The following constitutes proof of this assertion.

Proof that as the standard deviation of a Log Normal distribution approaches 0 the distribution becomes a Normal distribution with a mean of e^u where u is the mean of the natural logarithms of the data set values.

1) Log Normal probability density function: pdf(x) = $\frac{1}{x\sqrt{2\pi\sigma^2}}$ $e^{\frac{-(\mathrm{Ln}(x)-u)^2}{2\sigma^2}}$; u = mean(ln(x)), σ = std_dev(ln(x))

2) Determine the mode of the Log Normal distribution i.e. $\frac{dy}{dx} = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{dy}{dx} \left(\frac{e^{-(\ln(x)-u)^2/2\sigma^2}}{x}\right) = 0$; solve for x

3) $\frac{dy}{dx} = e^{-(\mathrm{Ln}(x)-u)^2/2\sigma^2} \left[\frac{-(\mathrm{Ln}(x)+u)}{\sigma^2} - 1\right] = 0$

4) Solve x for $\frac{-\mathrm{Ln}(x)+u}{\sigma^2} - 1 = 0$

5) $\mathrm{Ln}(x) = u-\sigma^2$

6) $x = e^{(u-\sigma^2)}$

7) As $\sigma \to 0$; $x \to e^u$

8) pdf(x) = $\frac{1}{x\sqrt{2\pi\sigma^2}} e^{\frac{-(\mathrm{Ln}(x)-u)^2}{2\sigma^2}}$

9) Taylor series of Ln(x) about e^u =

10) $\mathrm{Ln}(e^u) + \frac{x-e^u}{e^u} - \frac{(x-e^u)^2}{2e^{2u}} + \frac{(x-e^u)^3}{3e^{3u}} - \frac{(x-e^u)^4}{4e^{4u}} + \ldots\ldots =$

11) $\ln(e^u) + \sum_{k=1}^{\infty} \frac{-(-1)^k (x - e^u)^k}{k e^{ku}}$

12) $\text{Ln}(x-e^u) \sim \text{Ln}(e^u) + \frac{x - e^u}{e^u}$ as $\sigma \to 0$

13) $\text{Ln}(x-e^u) \sim u + \frac{x - e^u}{e^u}$

14) $\text{pdf}(x) \sim \frac{1}{x\sqrt{2\pi\sigma^2}} e^{\frac{-(u + \frac{x-e^u}{e^u} - u)^2}{2\sigma^2}}$

15) $\text{pdf}(x) = \sim \frac{1}{e^u\sqrt{2\pi\sigma^2}} e^{\frac{-(\frac{x-e^u}{e^u})^2}{2\sigma^2}}$ as $\sigma \to 0$

16) $\text{pdf}(x) \sim \frac{1}{\sqrt{2\pi(\sigma e^u)^2}} e^{\frac{-(x-e^u)^2}{2(\sigma e^u)^2}}$

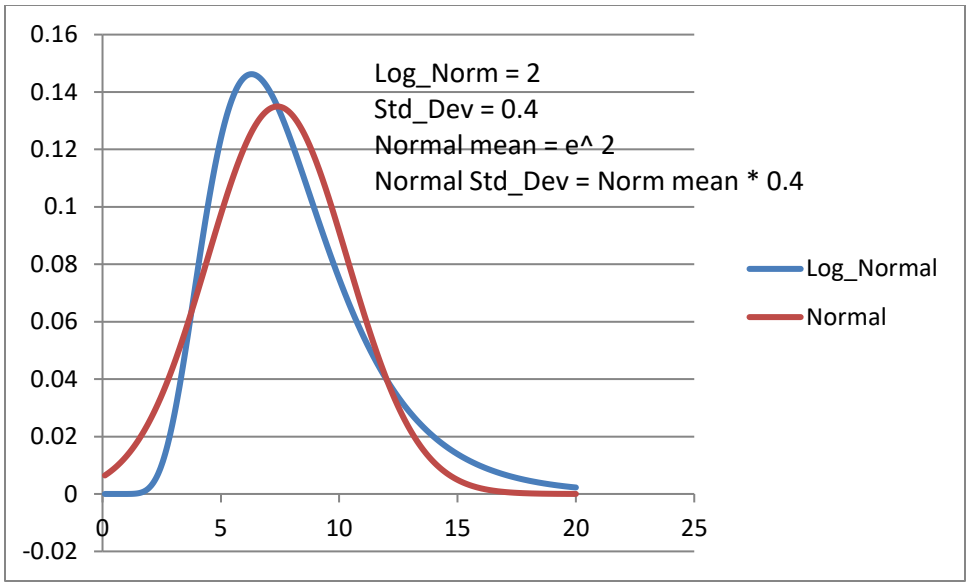
17) $u_x = \text{mean}(x)$; $\boldsymbol{\sigma}_x = \text{std\_dev}(x)$

18) $u_x \sim e^u$; $\boldsymbol{\sigma}_x \sim u_x \sigma$

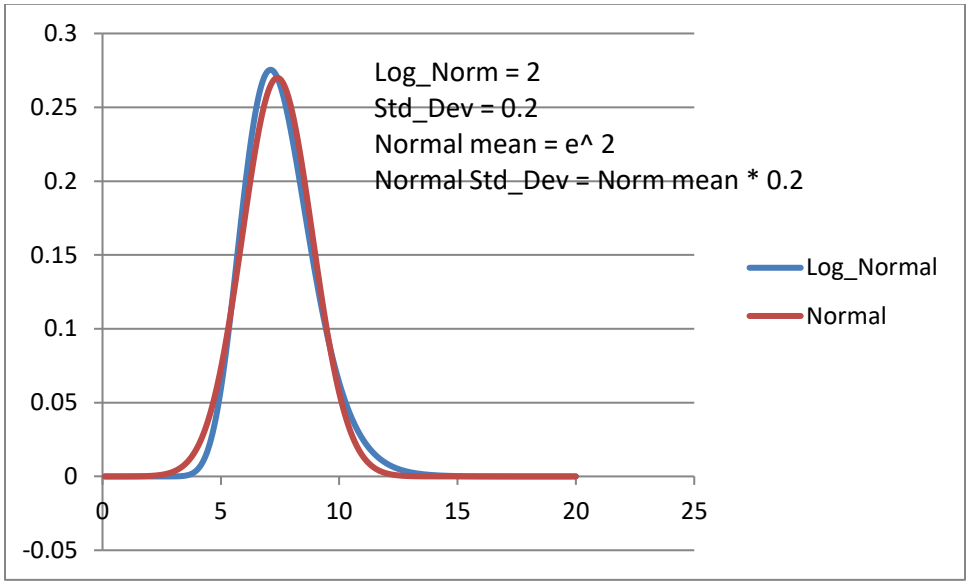
19) $\text{pdf}(x) \sim \frac{1}{\sqrt{2\pi(\sigma_x)^2}} e^{\frac{-(x-u_x)^2}{2(\sigma_x)^2}}$

20) Which is a Normal Distribution with a mean of e^u and a standard deviation of σe^u

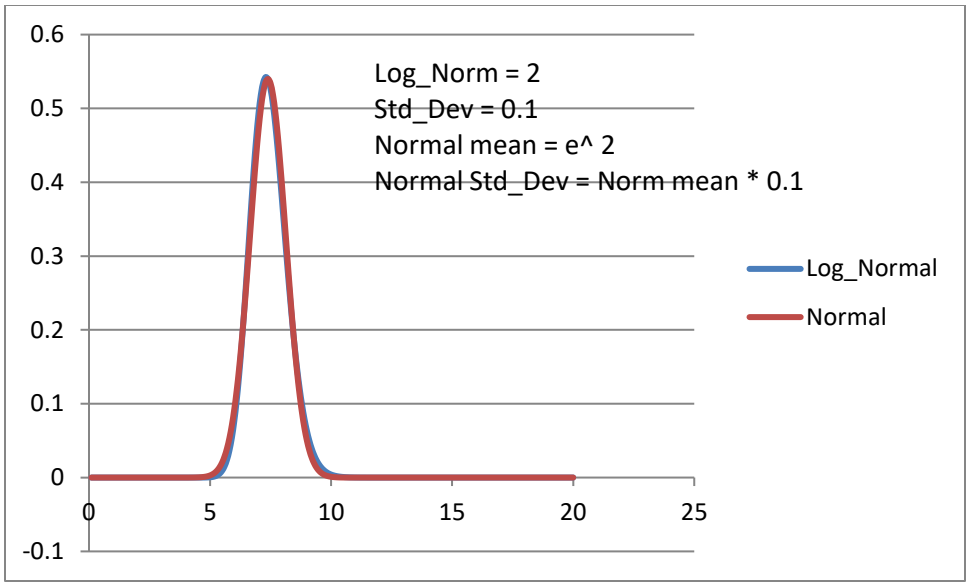
The following graphs are plots of the Log Normal distribution with given values of mean (u) and standard deviation of σ v. the Normal distribution with a mean of e^u and a standard deviation of σe^u.


0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
-0.02
0
5
10
15
20
25
Log_Norm = 2
Std_Dev = 0.4
Normal mean = e^ 2
Normal Std_Dev = Norm mean * 0.4
Log_Normal
Normal

Fig#4


0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
0
5
10
15
20
25
Log_Norm = 2
Std_Dev = 0.2
Normal mean = e^ 2
Normal Std_Dev = Norm mean * 0.2
Log_Normal
Normal

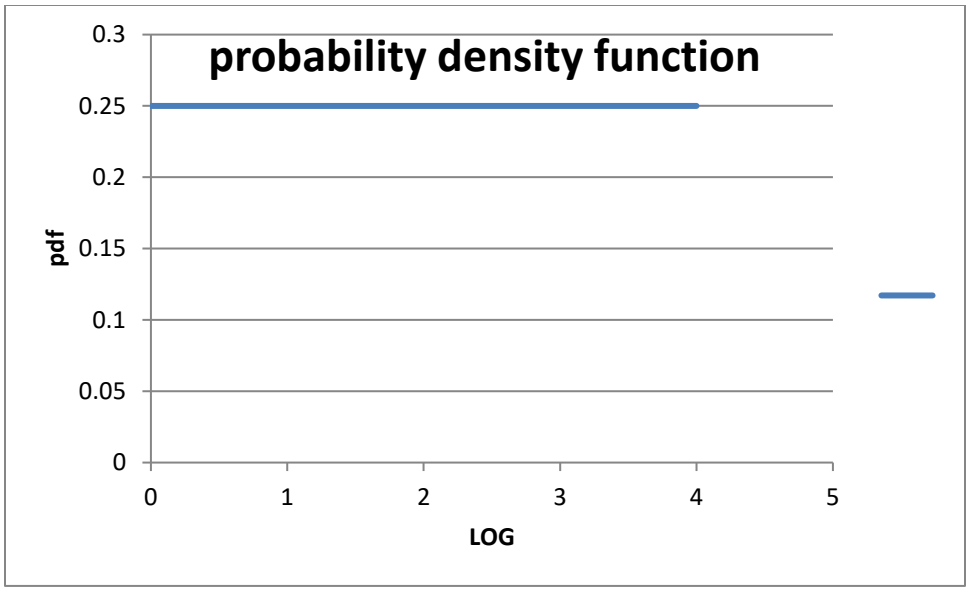
Fig# 5



Fig#6

The plots were derived from a Microsoft Excel spreadsheet. They strongly indicate empirically that as the standard deviation does approach zero the Log Normal distribution with a given mean and standard deviation does converge to a Normal distribution with a mean equal to e raised to the Log Normal mean and the standard deviation equal to the Normal mean times the Log Normal standard deviation.

The logarithm of the data generated from the probability density function of an exponential function is uniform throughout all orders of magnitude.



Fig#7-example of a probability density function of an exponential function

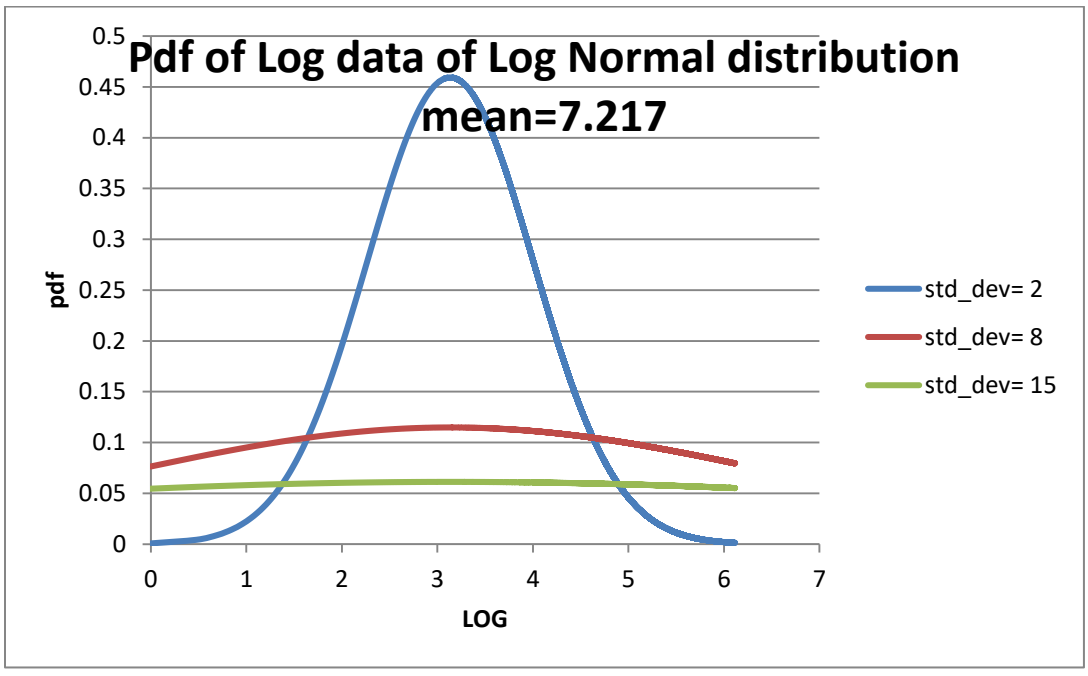
The pdf of the logarithm of data generated from a Log Normal distribution and most other probability density functions, such as gamma, Chi Square, Weibull begin and end on the x-axis unlike the exponential function, which is a uniform straight line. For instance, the pdf of the logarithm of a data set generated from a Log Normal distribution is a Normal distribution with respect to Lnx or Log(x).

The following argument constitutes a proof that the probability density function of the logarithm of data that conforms to a Log Normal distribution is a Gaussian or Normal distribution.

1. For a Lognormal distribution the pdf_x (probability density function) = $\frac{1}{x\sqrt{2\pi\sigma^2}} e^{-(\mathrm{Ln}(x)-u)^2/2\sigma^2}$
2. $Y = \mathrm{Log}_{10}(x)$

3. $pdf_y\, dy = pdf_x dx$
4. $pdf_y = pdf_x \frac{dx}{dy}$
5. $\frac{dy}{dx} = \frac{1}{xLn(10)}$; $\frac{dx}{dy} = xLn(10)$
6. $x = 10^y$
7. $pdf_y(Log(x)) = Ln(10) * 10^{\log(x)} \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\left(Ln\left(10^{\log(x)}\right)-u\right)^2/2\sigma^2} =$
8. $(x)*Ln(10) \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-(Ln(x)-u)^2/2\sigma^2} = Ln(10) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(Ln(x)-u)^2/2\sigma^2}$, which is a Gaussian distribution with respect to log(x)

Figures 6-8 Illustrate the probability density function of the logarithm of a data set that conforms to a Lognormal distribution and how it approaches a uniform distribution of a true Benford distribution as the Standard deviation increases.



Fig#8 – Probability Density Function of the Logarithm of a Data Set that Conforms to a Lognormal Distribution

For an exponential distribution, the mantissas between integral powers of ten (IPOT) are uniform since the probability density function is 1. This accounts for the fact that numbers beginning with 1 occur about 30% of the time and numbers beginning with 9 occur about 4.6% of the time.

For a Lognormal distribution or any other distribution if it can be shown that the probability density function of the sum of each mantissa for each corresponding IPOT approaches a constant value as the number of number of integral powers of ten (IPOT) approaches infinity the data set will conform to Benford's Law. The following argument constitutes a proof of this assertion.

Proof that if the probability density function of the logarithm of a data set is continuous and begins and ends on the x-axis and the number of integral power of ten (IPOT) values approaches infinity then the probability density function of the resulting mantissas will be uniform and; therefore, the data set will conform to Benford's law

1) The probability density function of a data set that conforms to Benford's Law is k/x = $\frac{1}{\ln(10)x}$
2) The probability density function of the log of the same function is a uniform distribution,
   a. pdf(y)dy = pdf(x)dx

b. $Y = \log(x) = \frac{\ln(x)}{\ln(10)}$

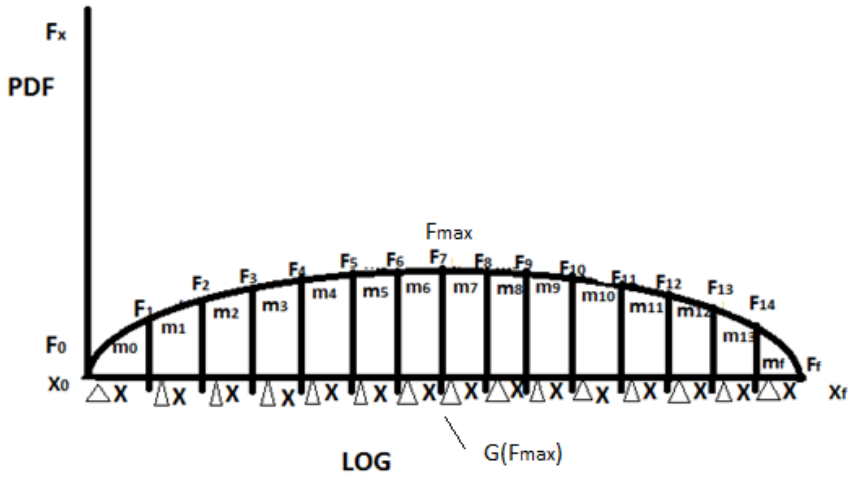
c. $pdf(y) = pdf(x)\frac{dx}{dy}$

d. $\frac{dy}{dx} = \frac{1}{x\ln(10)}$

e. $\frac{dx}{dy} = x\ln(10)$

f. $pdf(y) = \frac{x\ln(10)}{x\ln(10)} = 1$ – Uniform Distribution

3) Therefore, If it can be shown that the pdf of the log of a function is uniform then the data set follows Benford's Law.



4) Y = F(x)

5) $Y' = \frac{d(F(x))}{dx}$

6) $\int_{Xo}^{Xf} Y' dx = \int_{Xo}^{Xf} F'(x) dx =$ F(Xf) – F(Xo) = 0

7) Avg Value of Y' = $\frac{1}{Xf-Xo} \int_{Xo}^{Xf} Y' dx = \frac{0}{Xf-Xo}$

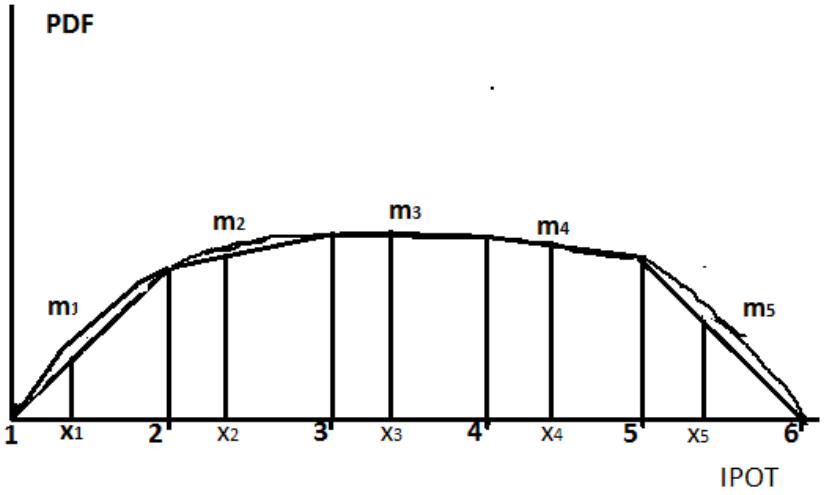
8) $F'_i(x) = \frac{F(i+1)-F(i)}{\Delta x}$; $\Delta x \rightarrow 0$

9) $\int_{Xo}^{Xf} F'(x) dx = 0$; $\sum_{i=0}^{N-1} \frac{F(i+1)-F(i)}{\Delta x} = 0$ as $\Delta X \rightarrow 0$

10) let m(i) = = $\frac{F(i+1)-F(i)}{\Delta x}$

11) $\sum_{i=0}^{N-1} m(i)\ \Delta X = 0$; $\Delta X \rightarrow 0$

Let's consider a simpler case.



12) Let $\Delta X = 1$

13) $m_1 + m_2 + m_3 + m_4 + m_5 = 0$

14) $\sum_{i=1}^{5} x_i = m_1x + m_1 + m_2x + m_1 + m_2 + m_3x + m_1 + m_2 + m_3 + m_4x +$

$m_1 + m_2 + m_3 + m_4 + m_5x = K$

15) $x(m_1 + m_2 + m_3 + m_4 + m_5) + m_1 + m_1 + m_1 + m_1 + m_2 + m_2 + m_2 + m_3 + m_3$

$+ m_4 = K$

16) $m_1 + m_2 + m_3 + m_4 + m_5 = 0$

17) $\sum_{i=1}^{5} x_i = 4m_1 + 3m_2 + 2m_3 + m_4 = K$ (constant)

18) AREA UNDER PDF = 1

18) $\int_1^6 f(x)\,dx = 1$

20) $\frac{m_1}{2} + m_1 + \frac{m_2}{2} + (m_1 + m_2) + \frac{m_3}{2} + (m_1 + m_2 + m_3) + \frac{m_4}{2} + (m_1 + m_2 + m_3 + m_4) + \frac{m_5}{2}$

$= 1$

21) $m_1 + m_2 + m_3 + m_4 + m_5 = 0$

22) $4m_1 + 3m_2 + 2m_3 + m_4 = 1$

Therefore K = 1

The sum of all functions at IPOT + x = 1 for any x.

The sum of all mantissas is a uniform distribution whose amplitude is equal to 1 and the PDF approaches a Benford distribution as $\frac{\Delta x}{n} \rightarrow 0$***.***

23) For the more general case:

24) $\sum_{i=1}^{r-1} m_i =$

25) $m_1x + m_2 + m_2x + m_1 + m_2 + m_3x + \ldots m_1 + m_2 + m_3 + \ldots m_{r-1}x =$

K

26) $x(m_1 + m_2 + \ldots + m_{r-1}) + (r-2)m_1 + (r-3)m_3 + \ldots + m_{r-2} = K$

27) $x(m_1 + m_2 + m_3 + m_{r-1}) = 0$

28) $(n-2)m_1 + (n-1)m_2 + \ldots + m_{r-2} = K$

29) $\frac{m_1}{2} + m_1 + \frac{m_2}{2} + m_1 + m_2 + \frac{m_3}{2} + m_1 + m_2 + m_3 + .. + m_{r-2} + \frac{m_{r-1}}{2} = K$

30) $\frac{1}{2}(m_1 + m_2 + m_3 + m_{r-1}) = 0$

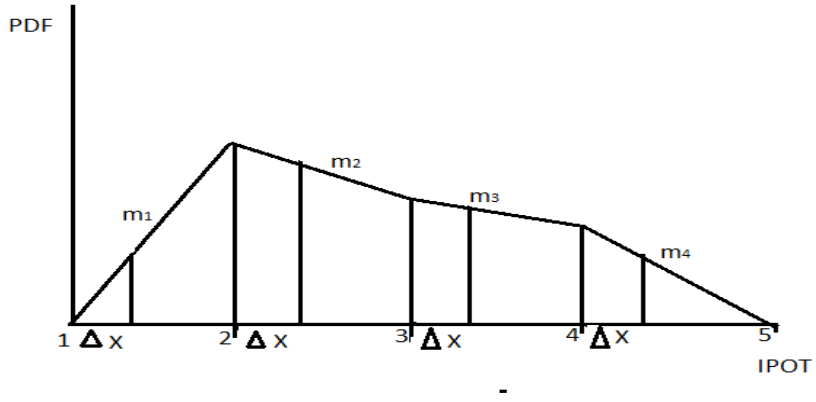
31) $(n-2)m_1 + (n-1)m_2 + ... + m_{r-2} = 1$

32) $K=1$

33) The sum of mantissa values at IPOT + x = 1 for any x

34) The sum of all mantissas is a uniform distribution whose amplitude is And, therefore, the PDF approaches a Benford distribution as $\frac{\Delta x}{N} \to 0$.

Proof that if the probability density function of the Logarithm a data set is continuous and begins and ends on the x-axis and the number of integral power of tens approaches infinity then the sum of probability distributions of all fixed intervals from all IPOT (ΔX) equals the interval itself (ΔX) itself.



-

1) $\sum_1^4 \int_i^{i+\Delta} \text{pdf}\, dx = \frac{1}{2} m_1(\Delta x)^2 + m_1 \Delta x + \frac{1}{2} m_2(\Delta x)^2 + (m_1 + m_2)\Delta x +$
$\frac{1}{2} m_3(\Delta x)^2 + (m_1 + m_2 + m_3)\Delta x + \frac{1}{2} m_4(\Delta x)^2 = K$

2) $\frac{1}{2}(\Delta x)^2 (m_1 + m_2 + m_3 + m_4) + (3m_1 + 2m_2 + m_3)\Delta x = K$

3) $m_1 + m_2 + m_3 + m_4 = 0$

4) $\frac{1}{2} m_1 + m_1 + + \frac{1}{2} m_2 + m_1 + m_2 + + \frac{1}{2} m_3 + m_1 + m_2 + m_3 + \frac{1}{2} m_4 =$

5) $\frac{1}{2}(m_1 + m_2 + m_3 + m_4) + 3 m_1 + 2 m_2 + m_3 = 1$

6) $3m_1 + 2m_2 + m_3 = 1$

7) $(3m_1 + 2m_2 + m_3)\Delta x = \Delta x$

8) $\sum_1^4 \int_i^{i+\Delta x} \text{pdf}\, dx = \Delta x$

In General:

9) $\sum_{i=1}^{r-1} \int_i^{i+\Delta x} \text{pdf}\, dx = \frac{1}{2}(\Delta x)^2 (m_1 + m_2 + m_3 + \ldots + m_{r-1}) +$

$$10) \qquad [(n-2)m_1 + (n-1)m_2 + \ldots + m_{r-2}]\Delta x = \Delta x$$

It can be easily shown that the fixed intervals don't have to start and end on an interval power of ten such as 10,100,1000 or 1,2,3 on a LOG plot as long as the fixed intervals are all offset by a power of ten.

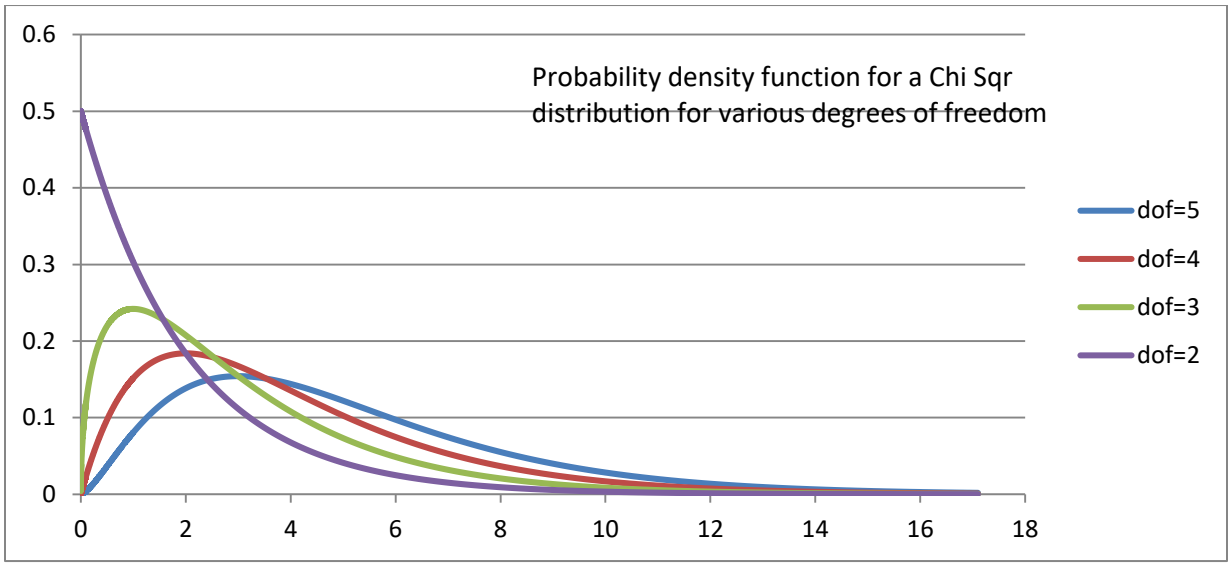
For instance, the left most interval starting point, where the curve intersects the x-axis, could be 2 with each succeeding interval 10 times the previous interval i.e 20,200,2000 etc. The data would still conform to Benford's Law with digit 1 contained in intervals 10-20, 100-200, 1000-2000; digit 2: 2-3,20-30,200-300;digit 3: 3-4,30-40,300-400;digit 4: 4-5,40-50,400-500;digit 5:5-6,50-60,500-600;digit 6:6-7,60-70,600-700;digit 7:7-8,70-80,700-800;digit 8:8-9,80-90,800-900;digit 9:9-10,90-100,900-1000. The first digit starts in the tens and ends in the 1000s; all of the others start in the single digits and end in the 100s. It's still the same result obtained by having the IPOT at each interval such as 1,10,100,1,000 etc.

This would explain why data sets that span many orders of magnitude conform very closely to Benford's law and data sets that span fewer orders of magnitude do not. This also explains why several other distributions such as gamma, beta, Weibull and exponential probability density functions conform fairly closely to **Benford's law and why Gaussian or Normal distributions do not (the pdf of the** logarithm of a Gaussian data span a very limited number of IPOTs. i.e.

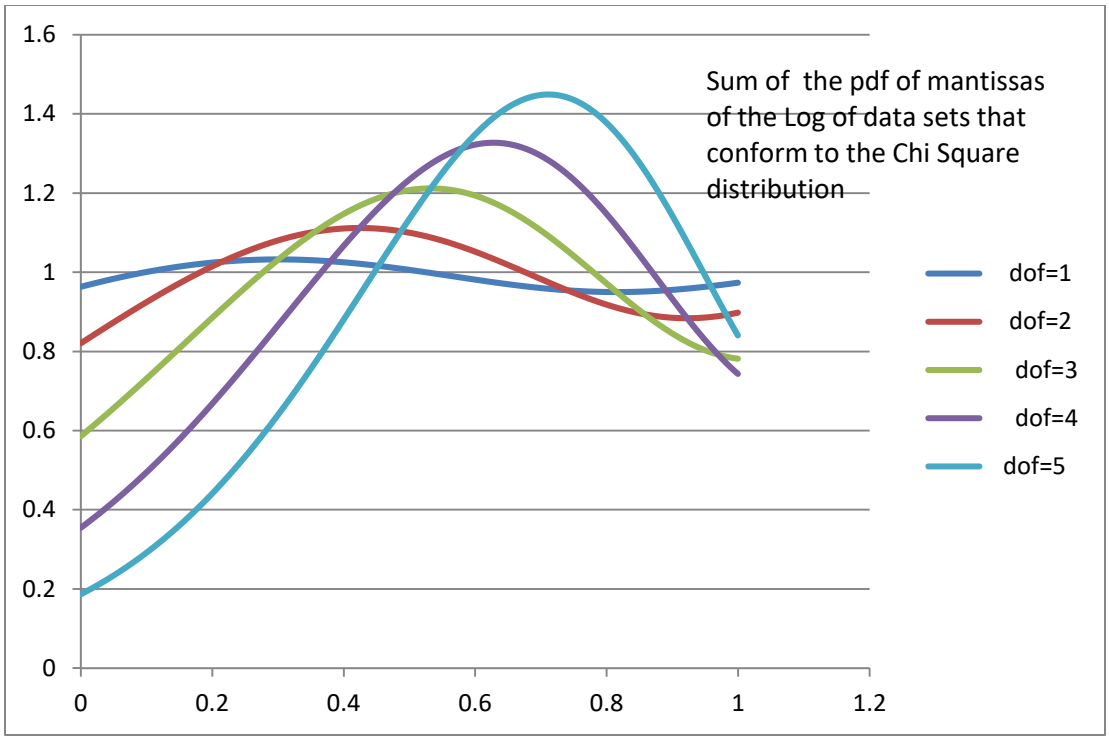
X* $\frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-(x-u)^2/2\sigma^2}$, the $e^{-(x-u)^2/2\sigma^2}$ term falls too rapidly.

More examples:

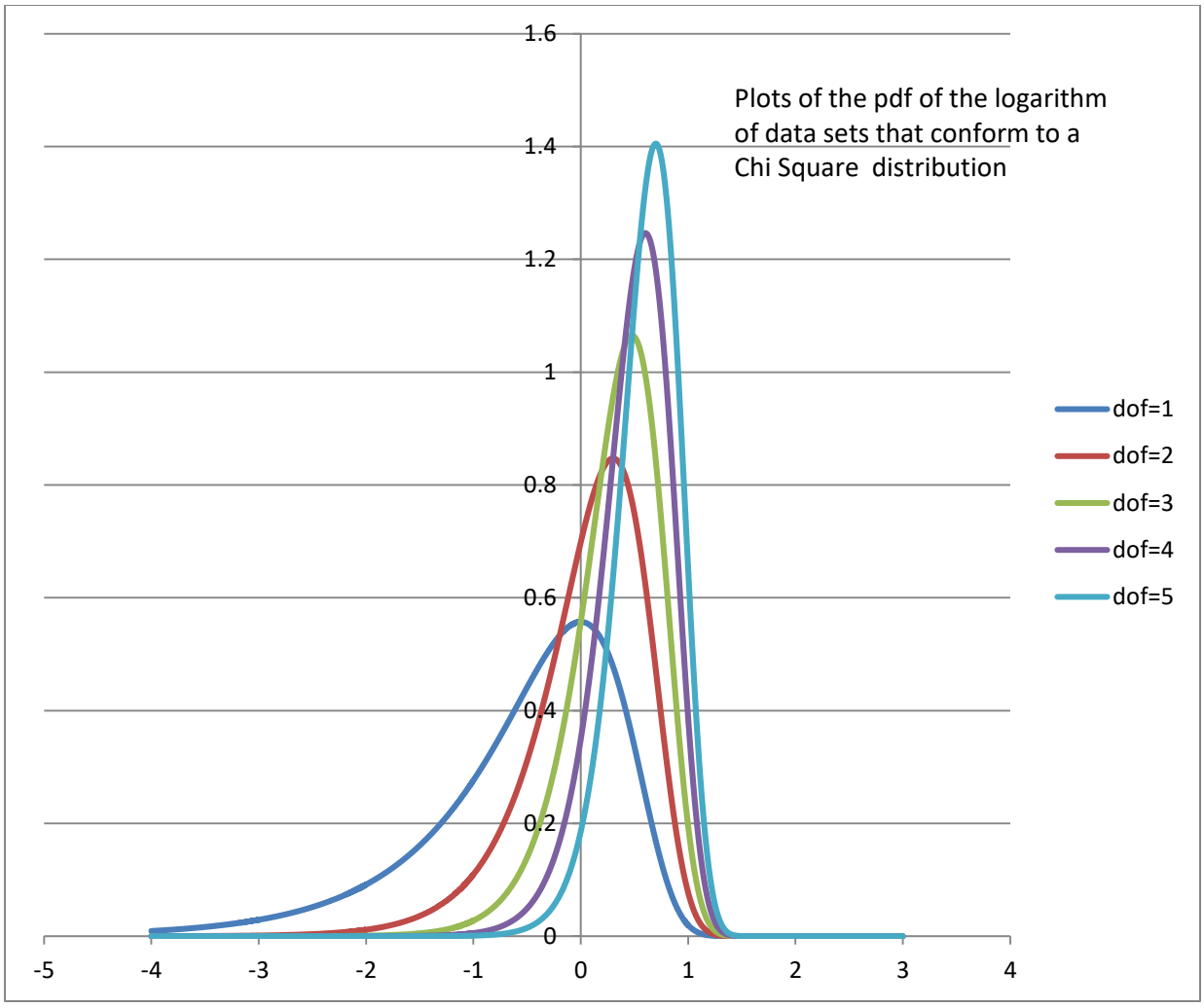
Chi Square distribution:



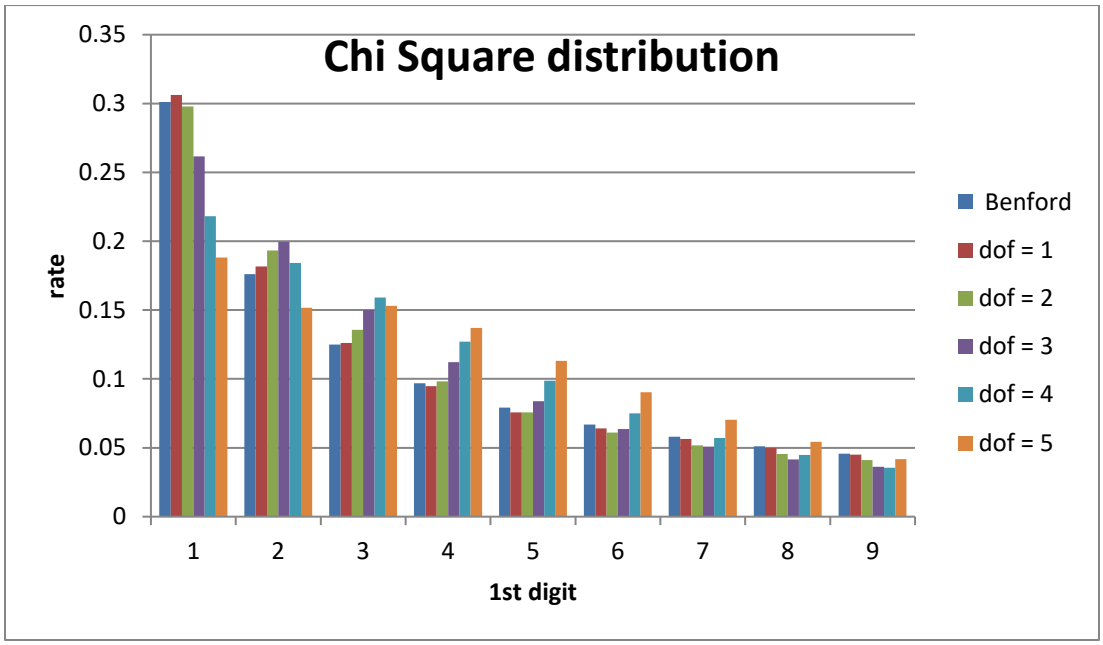
Fig#9- Probability density functions of a Chi Square distribution for various degrees of freedom



Fig#10 - Sum of the pdf of each mantissa for each corresponding IPOT for the Logarithm of data sets that conform to the Chi Square distribution

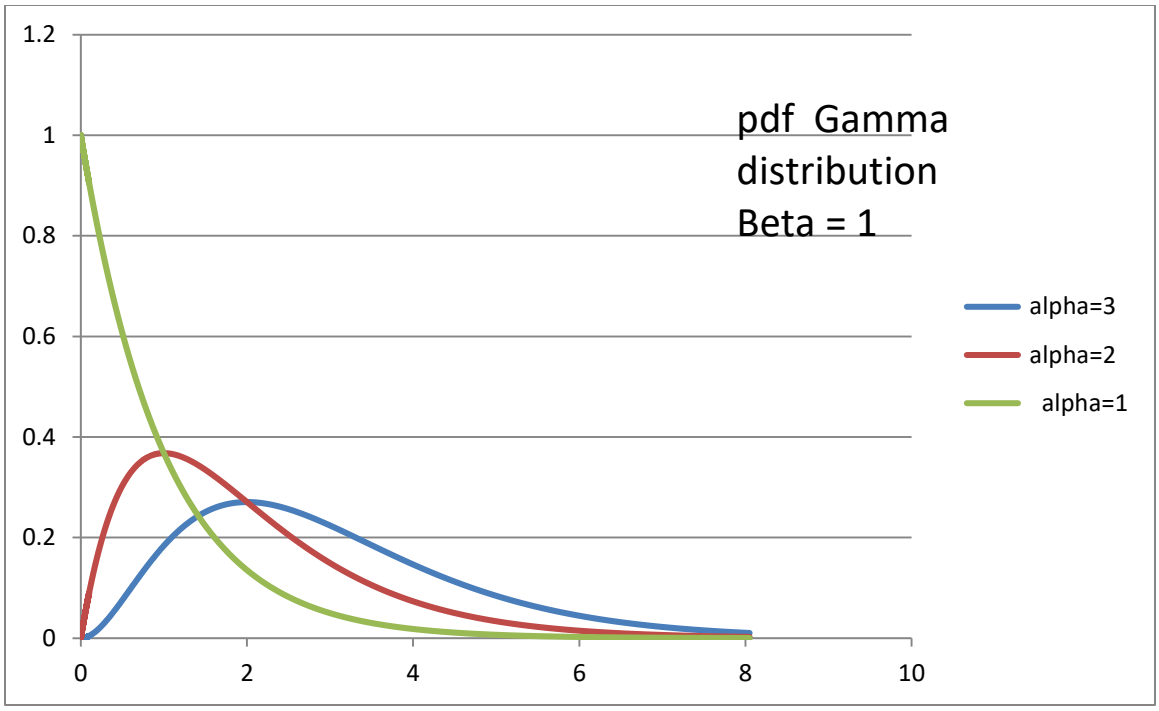


Fig#11- Plots of the probability density functions of the logarithm of data sets that conform to a Chi Square distribution

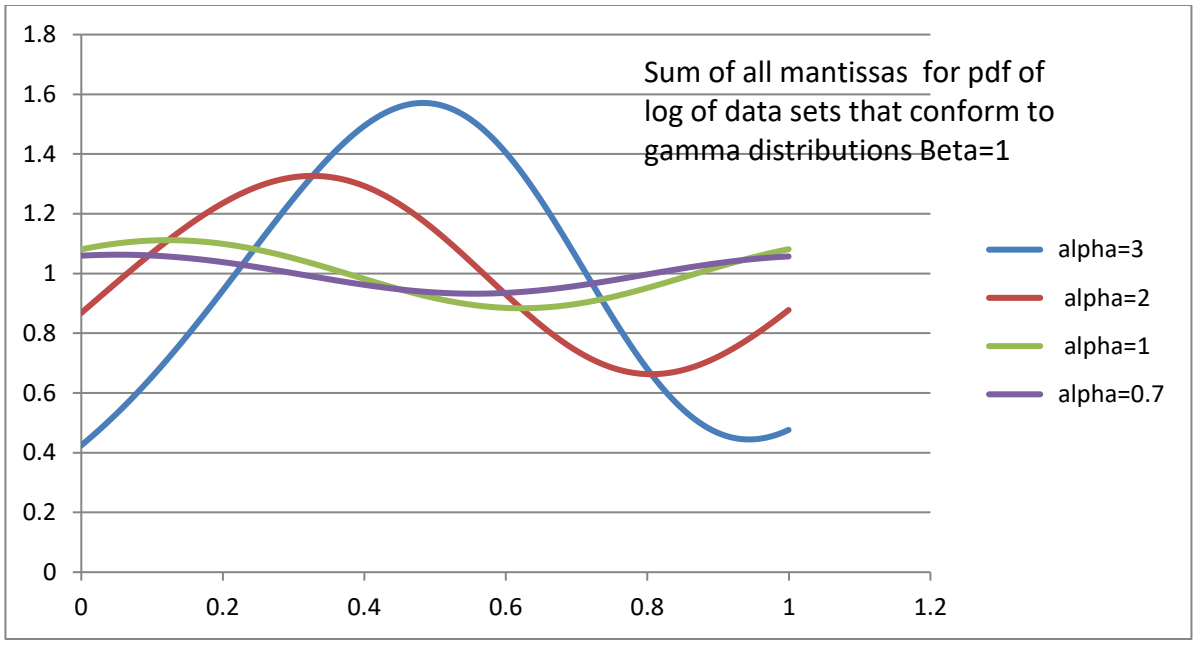


Fig#12 - 1st digit distribution for a Chi Square distribution for various degrees of freedom

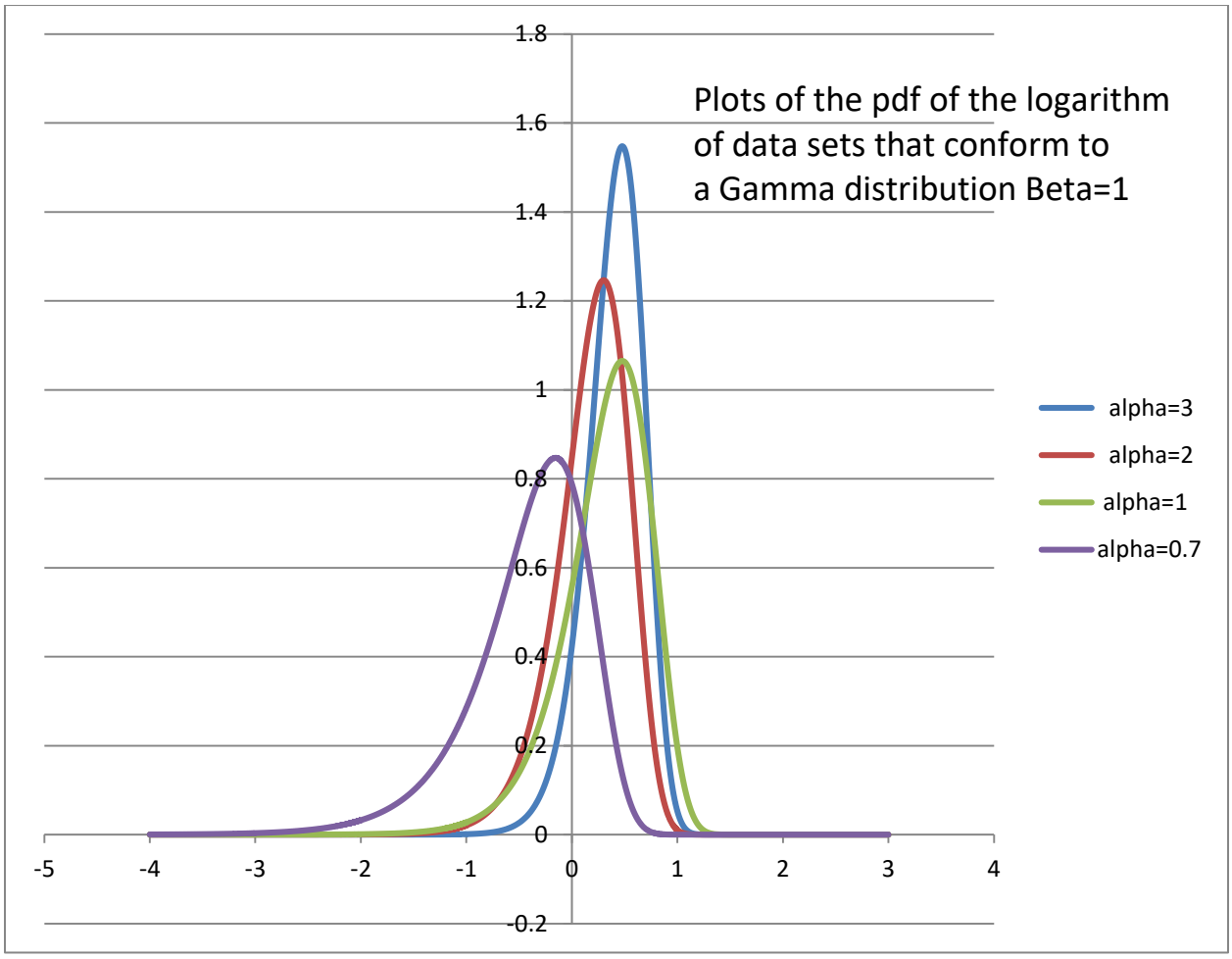
Gamma distribution:



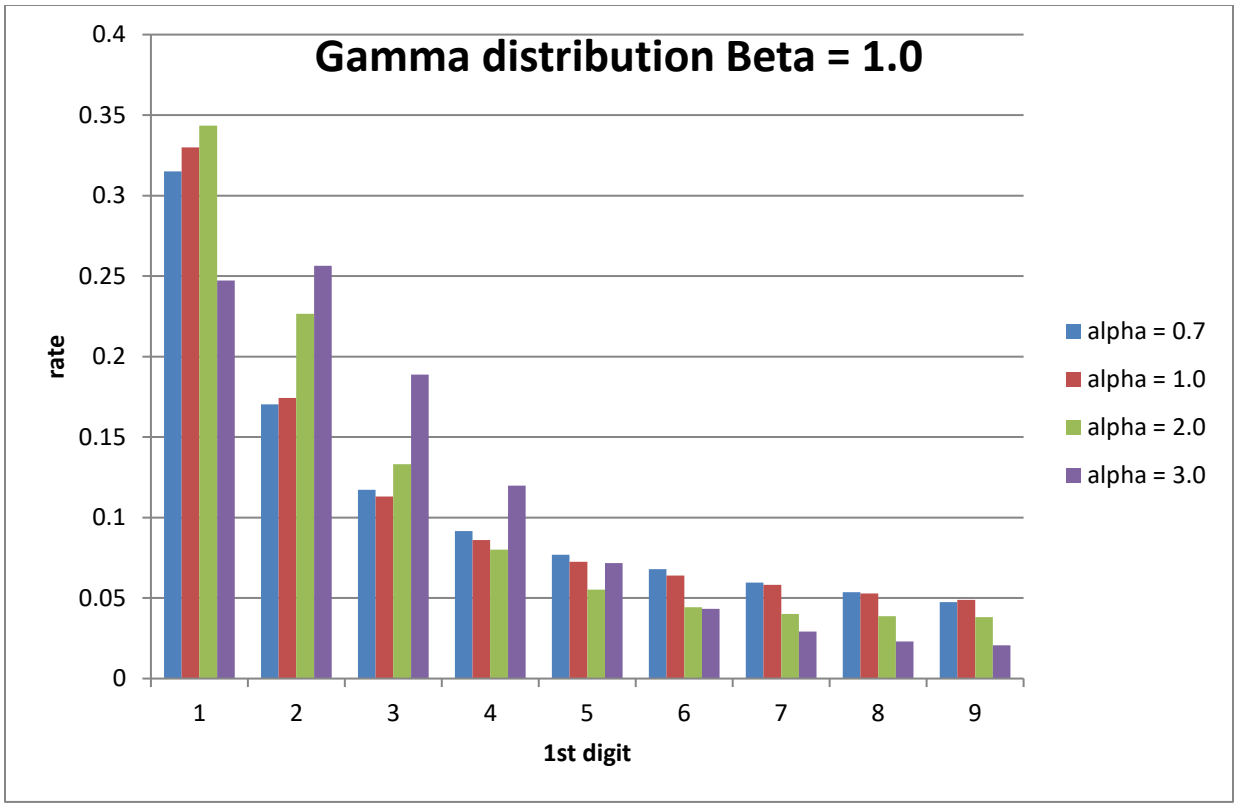
Fig#13 - Probability density function of a Gamma function Beta = 1 and various values for alpha



Fig#14 - Sum of the pdf of each mantissa for each corresponding IPOT of the logarithm of data sets that conform to Gamma distribution for Beta = 1



Fig#15- Plots of the probability density functions of the logarithm of data sets that conform to a Gamma distribution



Fig#16 - 1st digit distribution for a Gamma distribution with various values of alpha

The Summation test:

The Summation test consists of adding all numbers that begin with a particular first digit or first two digits and determining its distribution with respect to these first or first two digits numbers. Most people familiar with this test believe that the distribution is a uniform distribution for any distribution that conforms to Benford's law i.e. the distribution of the mantissas of the logarithm of the data set is uniform U[0,1). This summation test that results in uniform distribution is true for an exponential function (geometric progression) i.e. $y = a^{kt}$ but not true for a data set that conforms to a Log Normal distribution even when the Log Normal distribution itself closely approximates Benford's Law.

When the summation test is applied to real data such as population of cities, time intervals between earthquakes, and financial data, which all closely conforms to Benford's law, the summation test results in a Benford like distribution and not a uniform distribution. Citing *Benford's Law,* page 273, author Dr Mark Nigrini, " The analysis included the summation test. For this test the sums are expected to be equal, but we have seen results where the summation test shows a Benford- like pattern for the sums." Citing *Benford's Law,* page 141, author Alex Kossovski, " Worse than the misapplication and confusion regarding the chi-sqr test, Summation Test stands out as one of the most misguided application in the whole field of Benford's Law, attaining recently the infamous status of a fictitious dogma and leading many accounting departments and tax authorities astray." He also states on page 145, "Indeed **all** summation tests on actual statistical and random data relating to accounting data and financial data, census data, single-issue physical data, and so forth, show a strong and

consistent bias towards higher sums for low digits, typically by a factor of 5 to 12 approximately in the competition between digit 1 and digit 9, there is not a single exception!”

The histograms of the logarithm of the aforementioned data tend to resemble a Normal distribution, which is the definition of a Log Normal distribution (the Central Limit theorem applied to random multiplications). Therefore, if it can be shown that the Summation test performed on data that conforms to a Log Normal distribution results in a Benford like distribution then the Summation test applied to most real world data that conforms to Benford’s law will also conform to a Benford like distribution and not a Uniform distribution.

The exponential case:

The probability density function of a purely exponential function is 1/xLn(10). The expected value of a data set within an interval a, b

Is = $\frac{\int_a^b x*\mathrm{pdf}\,dx}{\int_a^b \mathrm{pdf}\,dx} = \frac{\frac{1}{\ln(10)}\int_a^b dx}{\frac{1}{\ln(10)}\int_a^b \frac{dx}{x}} = \frac{b-a}{\ln\frac{b}{a}}$

The sum of numbers within an interval a, b = expected value within an Interval a, b * the number of data points within the same interval

The number of data points within an interval a, b = N (total number of data points) * $\int_a^b \mathrm{pdf}\,dx = \frac{N*\frac{1}{\ln(10)}\int_a^b \frac{dx}{x}}{\frac{1}{\ln(10)}\int_1^{10} \frac{dx}{x}} = \frac{\ln(\frac{b}{a})}{\ln(10)}$ contained within an integral power of ten, $(10^k,10^{k+1})$

Therefore the sum is: $\frac{\mathrm{b}-\mathrm{a}}{\ln(\frac{\mathrm{b}}{\mathrm{a}})} * \frac{N l n(\frac{\mathrm{b}}{\mathrm{a}})}{\ln(10)} = \frac{N(\mathrm{b}-\mathrm{a})}{\ln(10)}$

Example: a=1, b=2; a=2, b=3 ….. a=9, b=10 Sum = $\frac{N}{\ln(10)}$

a=10, b=20 ……. a=90, b=100 Sum = $\frac{10N}{\ln(10)}$

Over several orders of magnitude =

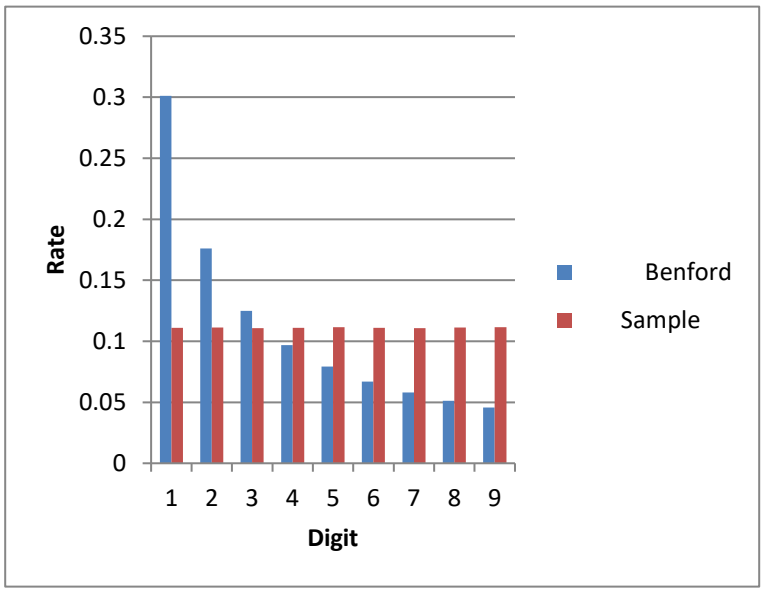
Sum = $\frac{N[b-a+10(b-a)+10^2(b-a)+10^3(b-a)+\cdots+10^k(b-a)]}{\ln(10)+\ln(10)+\ln(10)+\ln(10)+\cdots+\ln(10)}$

Generally*: Sum = $\frac{N}{\log_{10}(\frac{\text{max value}}{\text{min value}})} * \frac{1}{\ln(10)} * \sum_{\log_{10}(\text{min value})}^{\log_{10}(\text{max value})-1} 10^{\mathrm{k}}$, b-a=1

*The assumption is made that the minimum and maximum are integral powers of 10 i.e. 1, 10, 100, etc.

Summation Test

Digit	Sample	Benford	Sample
1	28931	0.301029996	0.111048
2	17082	0.176091259	0.1112
3	11764	0.124938737	0.110844
4	9424	0.096910013	0.110959
5	7520	0.079181246	0.111414
6	6507	0.06694679	0.110971
7	5588	0.057991947	0.11068
8	4977	0.051152522	0.111238
9	4428	0.045757491	0.111646
Total	96221		



Fig# -17 Summation with Respect to the 1st Digits i.e. 1,2,3,4,5,6,7,8,9 of an Exponential Function

Proof that the probability distribution of the sum of the values of a Log Normal probability density function with respect to the first digits (1 through 9) is nearly a Benford distribution and not a uniform distribution.

The probability distribution function of the sum of the values of a Benford probability density function (1/x) with respect to the first digits is a uniform distribution but such is not the case for a Lognormal density function.

Most numbers encountered in real life such as populations, scientific data, and accounting data are derived from the multiplication of statistically independent numbers, which constitute a Lognormal probability density function analogous to a Gaussian or Normal probability density function, which is derived from the addition of statistically independent numbers.

The following argument constitutes a proof that the sum of these numbers with respect to the first digits is nearly a Benford distribution as well as the number of values with respect to the first digits.

1. Pdf_x (probability density function) = f(x)
2. Average value = $\frac{\int_a^b xf(x)\,dx}{\int_a^b f(x)dx}$
3. Number of samples between a and b = $N\int_a^b f(x)dx$
4. Sum of values between a and b is Average value X number of samples between a and b =
5. = $\frac{\int_a^b xf(x)\,dx}{\int_a^b f(x)dx}$ X $N\int_a^b f(x)dx$ =
6. $N\int_a^b xf(x)dx$
7. For Lognormal distribution f(x) = $\frac{e^{-(\ln(x)-u)^2/2\sigma^2}}{x\sqrt{2\pi\sigma^2}}$
8. $N\int_a^b xf(x)dx$ = $N\int_a^b \frac{e^{-(\ln(x)-u)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$ dx
9. Assume Pdf_x = $\frac{e^{-(\ln(x)-u)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$

10. y = Log(x)
11. $Pdf_y\ dy = Pdf_x\ dx$
12. $Pdf_y = Pdf_x \frac{dx}{dy}$
13. $\frac{dy}{dx} = \frac{1}{x\ln(10)}$; $\frac{dx}{dy} = x \ln(10)$
14. If Log plot of can be $Pdf_y (\log(x)) = \ln(10)10^{\log(x)} \frac{e^{-\left(\ln(10^{\log(x)})-u\right)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} =$
15. $(x)* \ln(10)* \frac{e^{-(\ln(x)-u)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$
16. approximated with straight line between Integral power of ten (IPOT) then
    Because the mantissa distribution approaches a uniform distribution the resulting distribution of
    The x will be a nearly Benford distribution.

17. Therefore, the Ist digit distribution of the sum of values should be a nearly Benford distribution
    Instead of a uniform distribution as previously thought.

Proof that the sum of numbers that conform to a Log Normal distribution and begin with a particular digit will approach a distribution conforming to Benford's Law and not a uniform distribution as the standard deviation of the Log Normal distribution approaches infinity

1. $Pdf_x (Log\_Normal) = \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{x\sqrt{2\pi\sigma^2}}$
2. Expected value = $\int_{-\infty}^{\infty} x * \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{x\sqrt{2\pi\sigma^2}} dx = \int_{-\infty}^{\infty} \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} dx = e^{m+\frac{\sigma^2}{2}}$

3. Expected value in interval a-b = $\frac{\int_a^b \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\,dx}{\int_a^b \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{x\sqrt{2\pi\sigma^2}}\,dx}$
4. Sum = Expected value * number of values within interval a-b
5. Number of values within interval a-b = N (total number of values) $* \int_a^b \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{x\sqrt{2\pi\sigma^2}}\,dx$
6. Sum = $\frac{\int_a^b \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\,dx}{\int_a^b \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{x\sqrt{2\pi\sigma^2}}\,dx} * N * \int_a^b \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{x\sqrt{2\pi\sigma^2}}\,dx =$ $N\int_a^b \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\,dx$
7. Let u = ln(x) – m; ln(x) = u + m; x = $e^{u+m} = e^u * e^m$; du = $\frac{dx}{x}$; dx = xdu
8. Sum = $N\int_{\ln(a)-m}^{\ln(b)-m} \frac{e^{-u^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} * \frac{e^u}{\sqrt{2\pi\sigma^2}} * \frac{e^m}{\sqrt{2\pi\sigma^2}}\,du = N\frac{e^m}{\sqrt{2\pi\sigma^2}} \int_{\ln(a)-m}^{\ln(b)-m} e^{\frac{-(u^2-2\sigma^2u)}{2\sigma^2}}\,du =$
9. $N\frac{e^m}{\sqrt{2\pi\sigma^2}} \int_{\ln(a)-m}^{\ln(b)-m} e^{\frac{-(u^2-2\sigma^2u+\sigma^4-\sigma^4)}{2\sigma^2}}\,du =$
10. $N\frac{e^m}{\sqrt{2\pi\sigma^2}} \int_{\ln(a)-m}^{\ln(b)-m} e^{\frac{-(u-\sigma^2)^2+\sigma^4}{2\sigma^2}}\,du = N\frac{e^m}{\sqrt{2\pi\sigma^2}} \int_{\ln(a)-m}^{\ln(b)-m} e^{\frac{-(u-\sigma^2)^2}{2\sigma^2}} * e^{\frac{\sigma^2}{2}}\,du =$
11. $N\frac{e^{m+\frac{\sigma^2}{2}}}{\sqrt{2\pi\sigma^2}} \int_{\ln(a)-m}^{\ln(b)-m} e^{\frac{-(u-\sigma^2)^2}{2\sigma^2}}\,du =$
12. $N\frac{e^m * e^{\sigma^2/2}}{\sqrt{2\pi\sigma^2}} \int_{\ln(a)-m}^{\ln(b)-m} e^{\frac{-(u-\sigma^2)^2}{2\sigma^2}}\,du$ as ln(a) → - ∞ and ln(b) → ∞ , Sum = N * $e^{m+\frac{\sigma^2}{2}}$

13. As $\sigma \to \infty$ Sum = $N \frac{e^m * e^{\sigma^2/2}}{\sqrt{2\pi\sigma^2}} \int_{\ln(a)-m}^{\ln(b)-m} e^{\frac{-\sigma^2}{2}} du = N \frac{e^m}{\sqrt{2\pi\sigma^2}}$ $\int_{\ln(a)-m}^{\ln(b)-m} du$ = $N \frac{e^m}{\sqrt{2\pi\sigma^2}}$*[ln(b) –m –(ln(a)-m)] = = $N \frac{e^m}{\sqrt{2\pi\sigma^2}}$*[ln(b) – ln(a)]

14. Let a=1; b=2 $N \frac{e^m}{\sqrt{2\pi\sigma^2}}$* ln(2)

15. Let a=1; b=10 $N \frac{e^m}{\sqrt{2\pi\sigma^2}}$* ln(10)

16. $\frac{N \frac{e^m}{\sqrt{2\pi\sigma^2}} * \ln(2)}{N \frac{e^m}{\sqrt{2\pi\sigma^2}} * \ln(10)}$ = $LOG_{10}(2)$

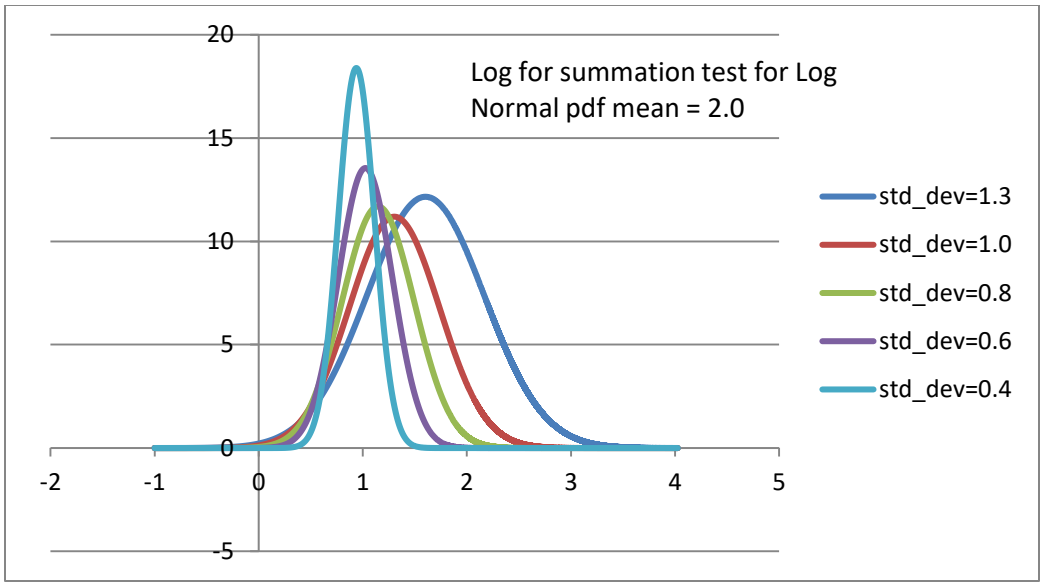
17. *Evaluated over all Integral powers of ten* =

18. $\frac{\int_{Ln(1)}^{\ln(2)} du + \int_{\ln(10)}^{\ln(20)} du + \int_{\ln(100)}^{\ln(200)} du + \cdots + \int_{\ln(1*10^k)}^{\ln(2*10^k)} du}{\int_{\ln(1)}^{\ln(10)} du + \int_{\ln(10)}^{\ln(100)} du + \int_{\ln(100)}^{\ln(1000)} du + \cdots + \int_{\ln(1*10^k)}^{\ln(2*10^{k+1})} du} = \frac{k*\ln(2)}{k*\ln(10)} =$ $LOG_{10}(2)$

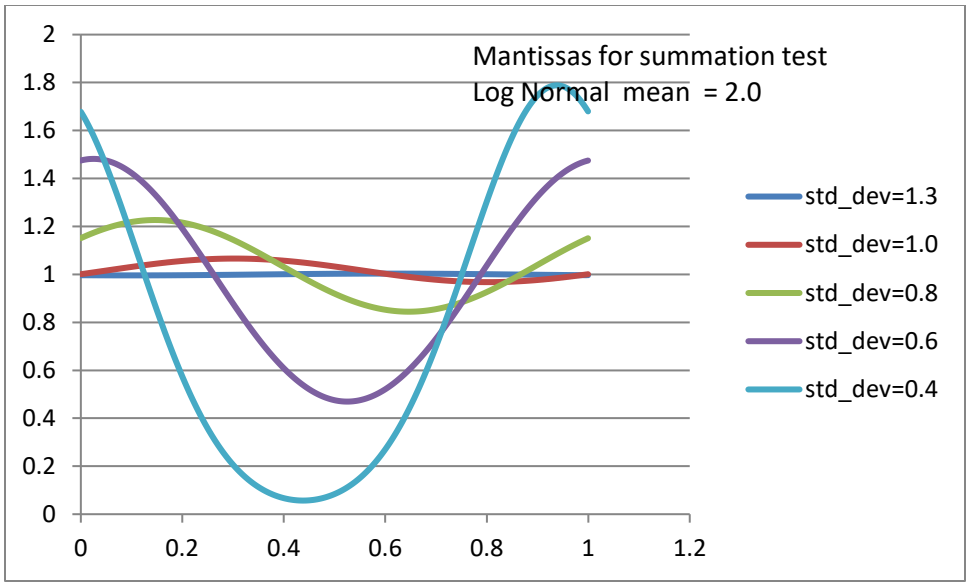
19. More Generally:

20. $= \frac{\int_{ln(d_1)}^{\ln(d_2)} du + \int_{\ln(d_10)}^{\ln(d_20)} du + \int_{\ln(d_100)}^{\ln(d_200)} du + \cdots + \int_{\ln(d_1*10^k)}^{\ln(d_2*10^k)} du}{\int_{\ln(1)}^{\ln(10)} du + \int_{\ln(10)}^{\ln(100)} du + \int_{\ln(100)}^{\ln(1000)} du + \cdots + \int_{\ln(1*10^k)}^{\ln(2*10^{k+1})} du} = \frac{k*\ln(\frac{d_2}{d_1})}{k*\ln(10)}$

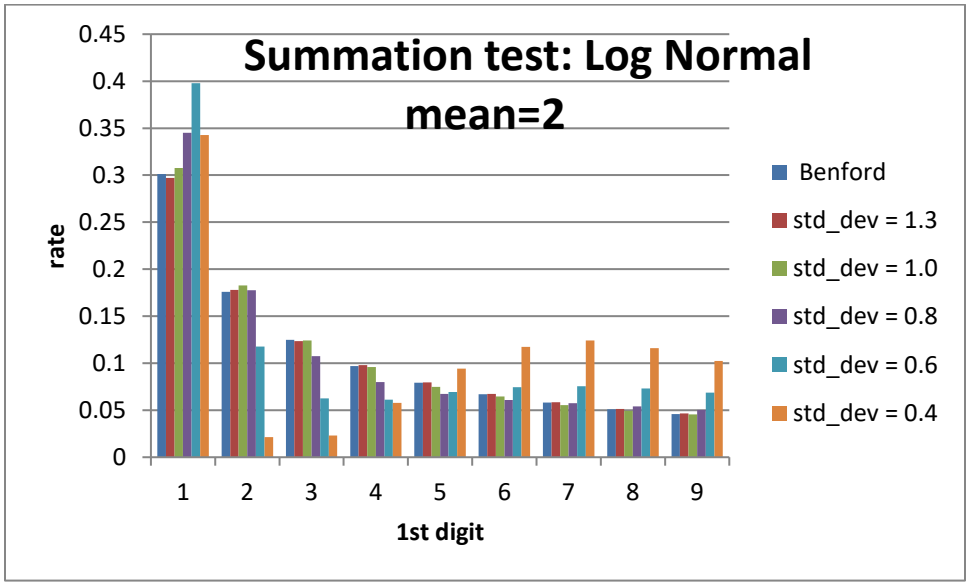
= $LOG_{10}(\frac{d_2}{d_1})$



Fig#17 - plot of the logarithm of the probability density function of the expected value (or sum) of a data set that conforms to a Log Normal distribution



Fig#18 - Sum of mantissas for Log Normal Summation test

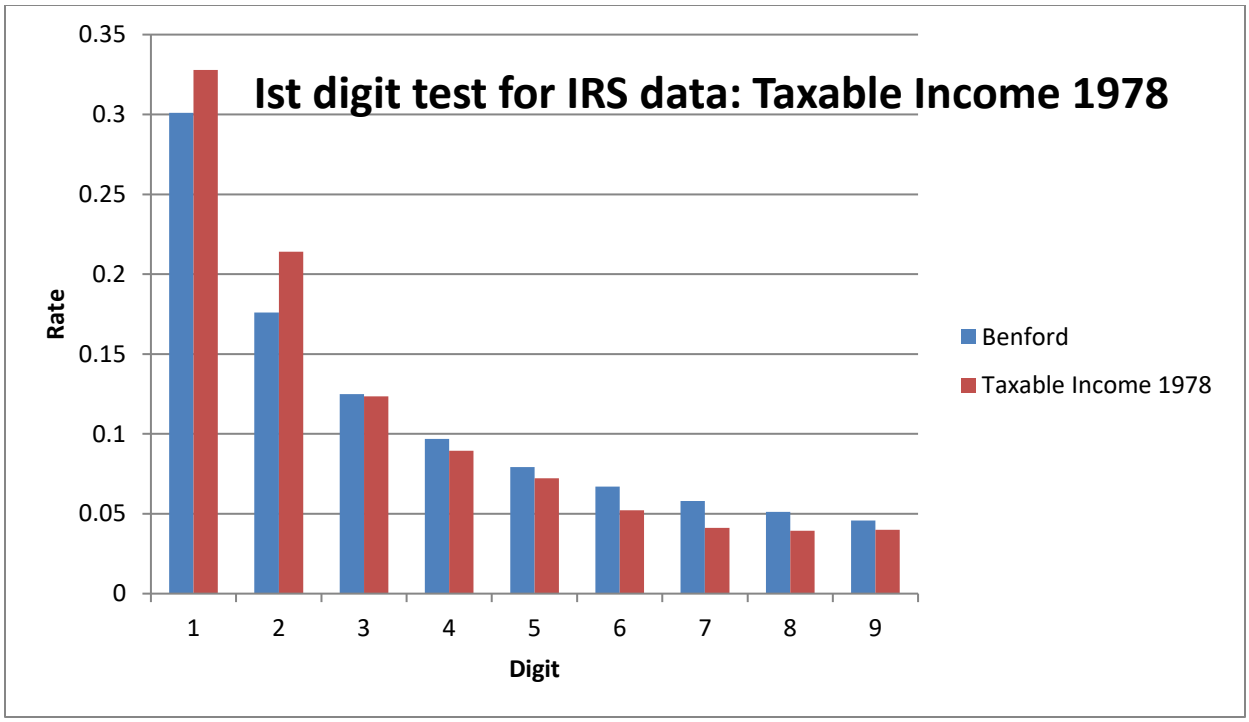


It is clear that the Summation test performed on a purely exponential function ($Y = a^{kt}$) results in a Uniform distribution. However, for data that conforms to a Log Normal distribution the Summation test in a Benford like distribution if the standard deviation is sufficiently large. This explains why the Summation test performed on a lot of real data

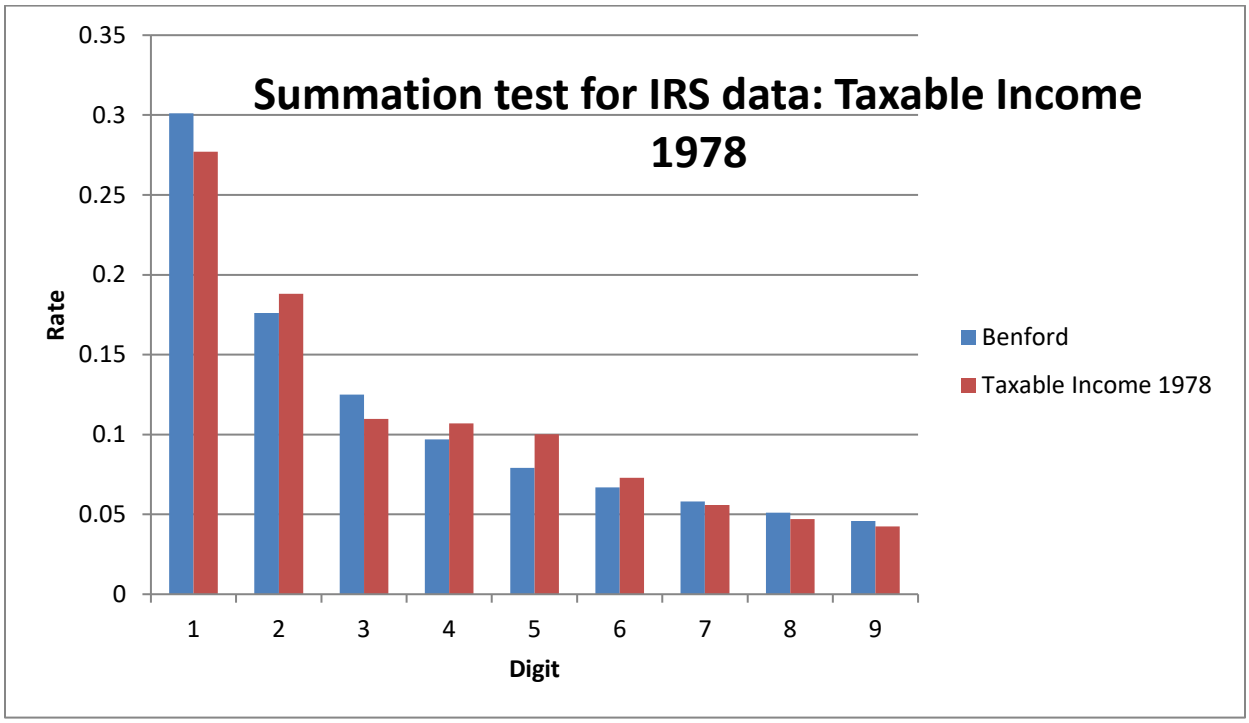
such as population, time interval between earthquakes, financial data results in a Benford like distribution, since the histograms closely resemble a Log Normal distribution

In recent years, Benford's law has been used to determine accounting and income tax fraud. Overall accounting and tax data should theoretically conform to Benford's law. The following analysis was conducted by me of taxable income data for the year of 1978. The results approximately follow Benford's law including the summation test, which is not what a lot of people conversant in Benford's would expect (it is not at all uniform). While examining the Ist digit test, it appears that there is a surplus of numbers that that begin with the lower numbers(1,2) and a dearth of numbers that begin with the higher digits(4,5,6,7,8,9) possibly indicating that people are understating their income (example: 9XXXX is reduced to 8XXXXX, 8XXXX is reduced to 7XXXXX and so on. The summation test could be used to determine if there are inordinate huge numbers(either one or many large numbers that begin with a particular digit).

Example: IRS data: Taxable Income – 1978



Fig#20 - Ist digit test for IRS data: Taxable Income 1978



Fig#21 - Summation test for IRS data: Taxable Income 1978

Conclusions:

1) The extent of conformity to Benford's law is to how closely the pdf of the mantissas of the logarithms of data set is uniform.
2) The more skewed the probability density function the more Benford like the distribution becomes, since $\int_{x1}^{x2} \mathrm{pdf}(x)dx = \int_{ax1}^{ax2} \mathrm{pdf}(x)dx$ (scale invariance) and if a is large then the pdf must be a much lower value for a considerable distance along the x axis
3) If the pdf of the logarithms of a data set begins and ends on the x axis and the curve between all integral power of tens (IPOT) can be approximated with a straight line the pdf will approach a Benford distribution.
4) The distribution of the logarithm of a data set is xpdf(x)
   a) If the data set is derived from an exponential function then the pdf of the logarithm of the data set will be uniform throughout all integral powers of ten.
   b) If the data set is derived from a Log Normal distribution then the pdf will be a Gaussian distribution will respect to lnx.
   c) Most distributions that involve an exponential component such as Gamma, Chi Square, Beta, Weibull will approach a Benford distribution as their respective standard deviations approach infinity, since xpdf(x) tends to start and end on the x axis over several integral powers of ten.
5) The summation test results in a uniform distribution for exponential functions; the pdf of the pdf xpdf(x) (expected values) is $\frac{x}{x\ln(10)} = \frac{1}{\ln(10)}$ for all ist digits. For Log Normal and the other aforementioned distributions the pdf of logarithm of the data set

ln(10)x^2pdf approaches a Benford distribution as the standard deviation approaches infinity.

References:

Berger, A and Hill, TP (2015), *An Introduction to Benford's Law,* **Princeton University Press: Princeton, NJ ISSN/ISBN 9780691163062**

Nigrini, MJ (2012), *Benford's Law: Applications for Forensic Accounting, Fraud Detection,* ***John Wiley and Sons,*** **ISSN/ISBN 978-1-118-15285-0**

Berger, A and Hill, TP(2010), *Fundamental Flaws in Feller's Classical Derivation of Benford,s Law (2010)***, Arxiv: 1005.2598v1**