Why the Summation Test Results in a Benford, and not a Uniform Distribution, for Data that Conforms to a Log Normal Distribution

R C Hall, MSEE, BSEE

e-mail: rhall20448@aol.com

Abstract

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). The summation test that results in a 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.

Introduction

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.

most as far as quantities are concerned. What happens on (10, 100) is by far less significant to sums of quantities.

Indeed **all** summation tests on actual statistical and random data relating to accounting 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!

In order to demonstrate the typical (uneven) behavior of these nine sums occurring in all random data, let us perform the summation test on the (almost perfectly) logarithmic U.S. Census data of Populations of Cities and Towns Incorporated. The largest population value of 8,391,881 belonging to New York City is considered an outlier, standing shoulder and above the rest of the other cities, and potentially swaying calculated sums a great deal. Hence NYC is being omitted altogether in this summation test. Figure 3.9 depicts the relevant chart showing clearly non-uniform sum series. In fact, this sum proportion does remind us a great deal of Benford's Law itself (except for the vertical scale)! Sums for digits 1 and 9 differ by a factor of 5.9, while in Benford's Law these proportions differ by a factor of 6.6, which is quite close. Clearly, summation test here did not

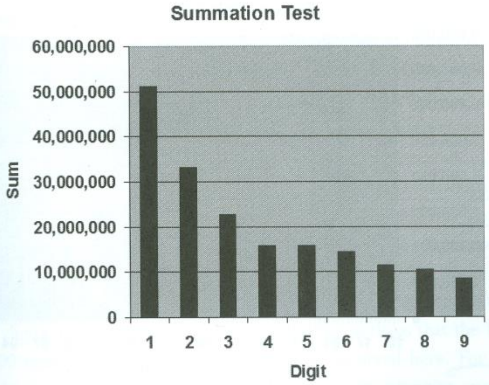


Figure 3.9 Sums Along First Digits, U.S. Populations Centers (New York City Excluded)

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

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

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

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

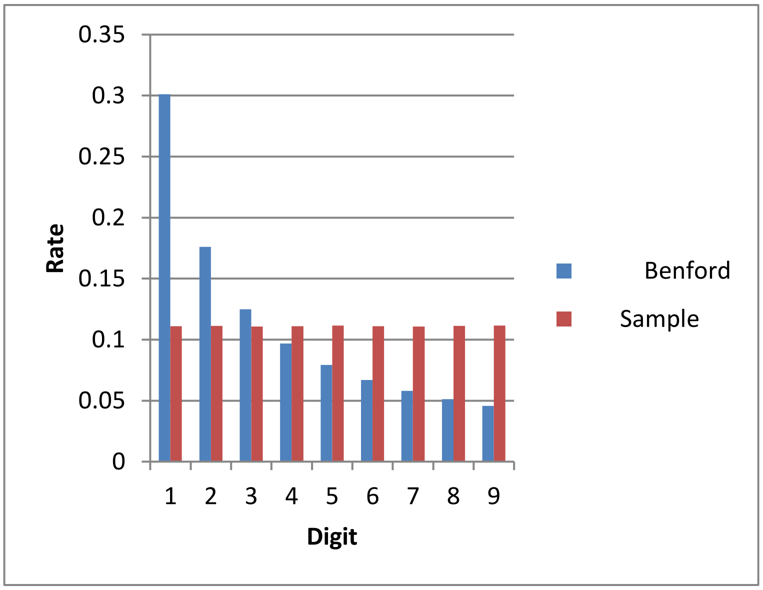
Generally*: Sum = $\frac{N}{log_{10}(\frac{max\ value}{\min value})}$ * $\frac{1}{\ln(10)}$ * $\sum_{log_{10}(\min value)}^{log_{10}(\max value)-1} 10^{\mathrm{k}}$, b-a=1

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

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

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		



The Log Normal case:

For the Log Normal distribution things are not quite the same. The data points themselves with respect to the numbers that begin with a particular digit will tend to conform to Benford's law as the standard deviation approaches infinity. The sum of all of the data points with respect to the first digits will also tend to conform to Benford's law (the distribution of the combined mantissas is uniform (U(0,1]).

The following argument constitutes a proof of this assertion.

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 = $\frac{\int_a^b \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\,\mathrm{dx}}{\int_a^b \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{x\sqrt{2\pi\sigma^2}}\,\mathrm{dx}}$
4. Sum = Expected value * number of values within interval a-b
5. Number of values within interval a-b = N (total number of values) * $\int_a^b \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{x\sqrt{2\pi\sigma^2}}\,\mathrm{dx}$

6. Sum = $\frac{\int_a^b \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\mathrm{dx}}{\int_a^b \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{x\sqrt{2\pi\sigma^2}}\mathrm{dx}}$ * N* $\int_a^b \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{x\sqrt{2\pi\sigma^2}}\mathrm{dx}$ = N$\int_a^b \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\mathrm{dx}$

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$\int_{\ln(a)-m}^{\ln(b)-m} \frac{e^{-u^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} * e^u * e^m$ 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 σ → ∞ 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. $\frac{\mathrm{N}\frac{e^m}{\sqrt{2\pi\sigma^2}} * \ln(2)}{\mathrm{N}\frac{e^m}{\sqrt{2\pi\sigma^2}} * \ln(10)}$ = $\mathrm{LOG}_{10}(2)$

17. $Evaluated\ over\ all\ Integral\ powers\ of\ ten$ =

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

19. More Generally:

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

See enclosure for an excel program that computes the distribution for the sums with respect to the first digits. The program essentially computes the equation listed in line 20 and does indicate a Benford distribution for most Log Normal distributions parameters that occur in the real world such as populations, financial data, time interval between earthquakes, etc.

Probability density function (pdf) for the logarithm of a data set

Given: pdf(x); x – data set

$Y = \log_{10}(x)$

1. $\mathrm{Pdf}_{\gamma}\, dy = \mathrm{Pdf}_{\chi}\, dx$
2. $\mathrm{Pdf}_{\gamma} = \mathrm{Pdf}_{\chi} \frac{dx}{dy}$
3. $dy = \frac{dx}{\ln(10)x}$
4. $\frac{dx}{dy} = x\ln(10) = 10^{y} \ln(10)$
5. $pdf_{\gamma} = 10^{y} \ln(10)\, \mathrm{Pdf}_{\chi}$
6. $pdf_{\gamma} = 10^{\log(x)} \ln(10)\, \mathrm{Pdf}_{\chi} = x\ln(10)\, \mathrm{Pdf}_{\chi}$

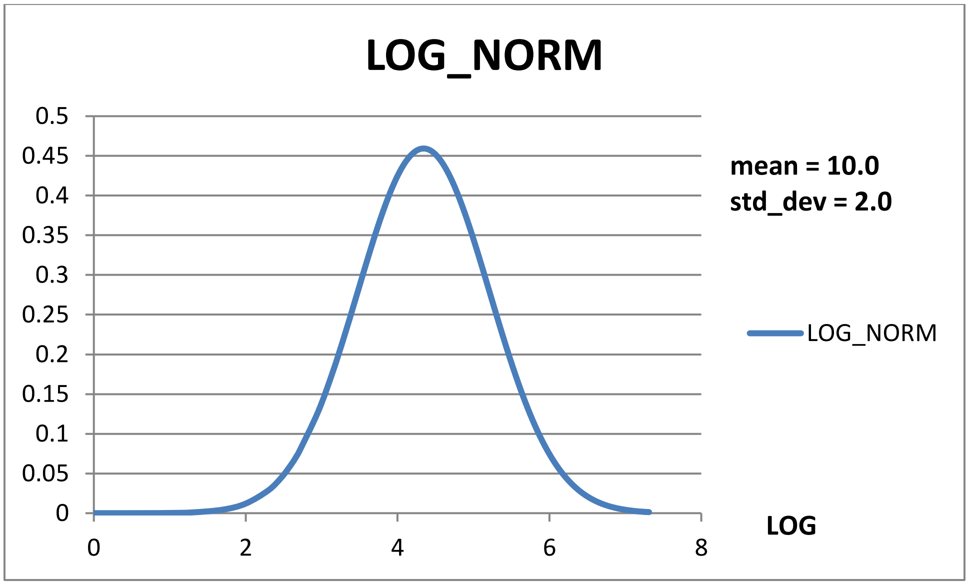
For a Log Normal distribution:) $\mathrm{Pdf}_{\chi} = \frac{e^{-(\ln(x)-u)^2/2\sigma^2}}{x\sqrt{2\pi\sigma^2}}$

$\mathrm{Pdf}_{\gamma} = x * \ln(10) * \frac{e^{-(\ln(x)-u)^2/2\sigma^2}}{x\sqrt{2\pi\sigma^2}} = \ln(10) * \frac{e^{-(\ln(x)-u)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$, which is a Gaussian or Normal distribution with respect to log x.

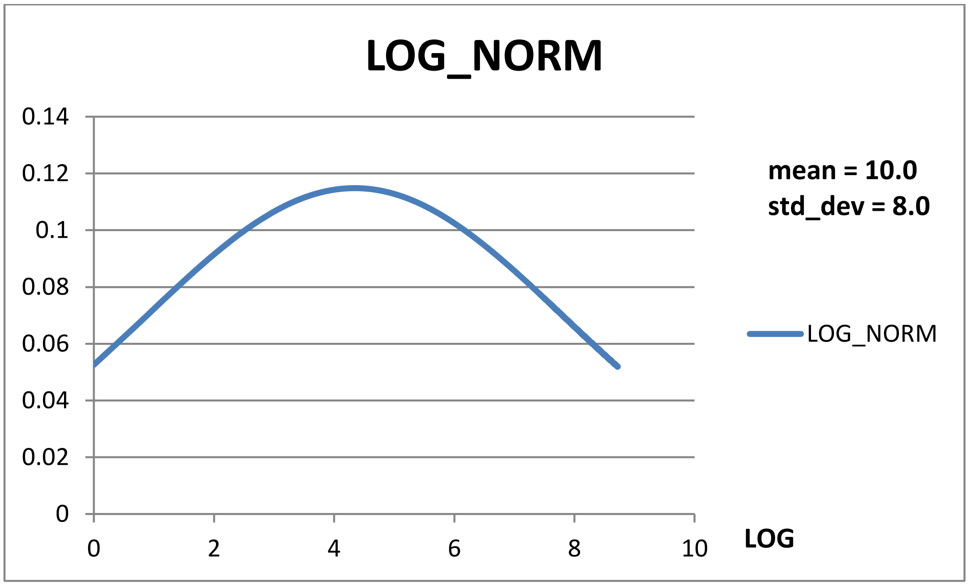
It can be shown that if the curvilinear distance between the integral powers of ten On the log plot can be approximated with a straight line then the distribution of the resultant mantissas will be a uniform distribution and, therefore, conform to a Benford distribution. See appendix A

I am aware of the apparent fallacy expressed in the paper " *Fundamental Flaws in Feller's Classical Derivation of Benford's Law* ", authors, Arno Berger and Theodore P. Hill that states " If the spread of a random variable X is very large, then X (mod 1) will be approximately uniformly distributed on [0,1].". I further stipulate that the function be continuous; start and end on the X-axis, and that the curvilinear distance between integral powers of ten can be approximated by a straight line. Also, the function does not have to start and end on an integral power of ten as long as the end value minus the start value is an integer and greater than two. This is tantamount to multiplying by a constant, which does not affect the outcome due to the scale invariance theorem as applied to Benford’s law.

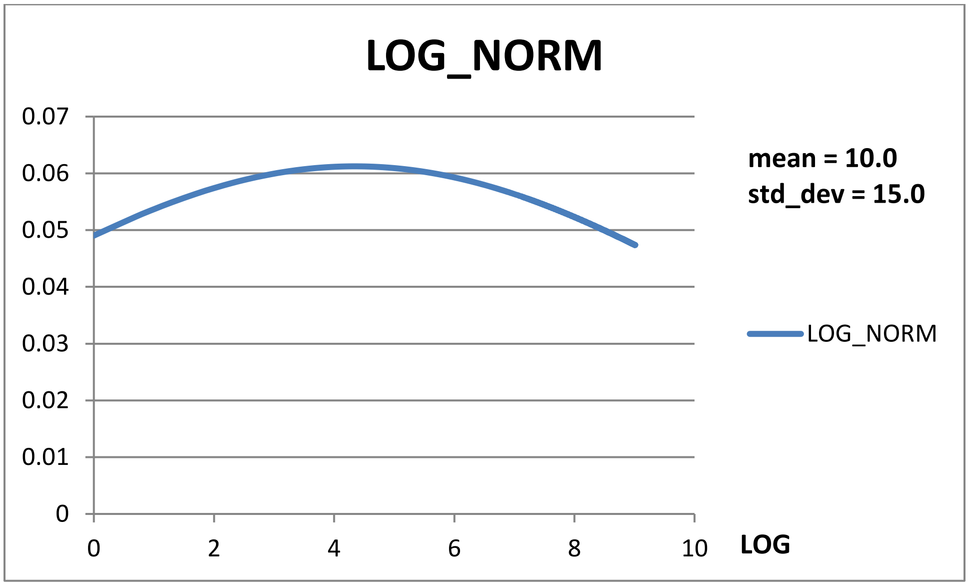
Fig#2 – Probability Density Function of the Logarithm of a Data Set that Conforms to a Log_Normal Distribution



Fig#3 – Probability Density Function of a Data Set that Conforms to a Log_Normal Distribution as the Standard Deviation Increases



Fig#4 - Probability Density Function of a Data Set that Conforms to a Log_Normal Distribution as the Standard Deviation Increases



One can observe that as the standard deviation increases the curvilinear distance between each interval power of ten approaches a straight line.

The sum of the data with respect to a given interval a, b is

N$\int_a^b \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$ dx

The probability density function (the integrand) is $\frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$

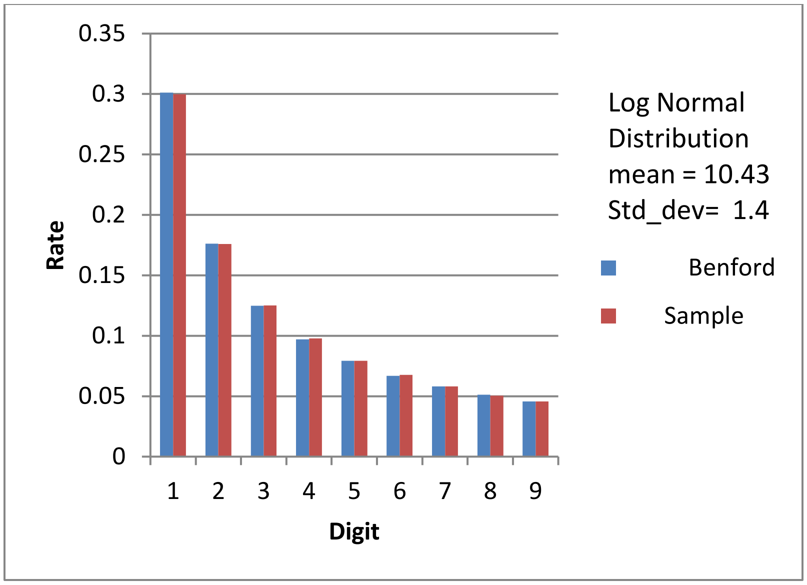
normalized by the mean.

.

The probability density function of logarithm of the data set is $x * Ln(10)\ \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$ normalized by the mean.

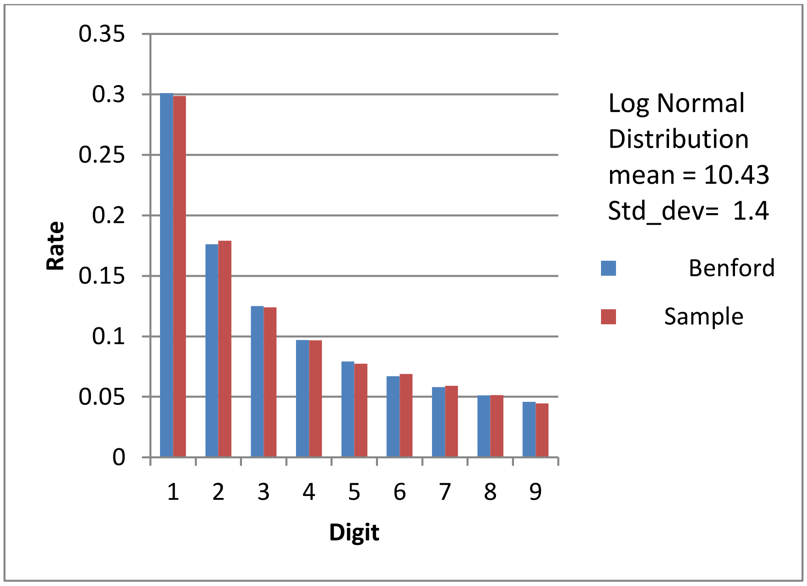
Again, if the curvilinear distance between the integral powers of ten can be approximated by a straight line then the distribution will approach a Benford distribution.

The following figures illustrate this phenomenon

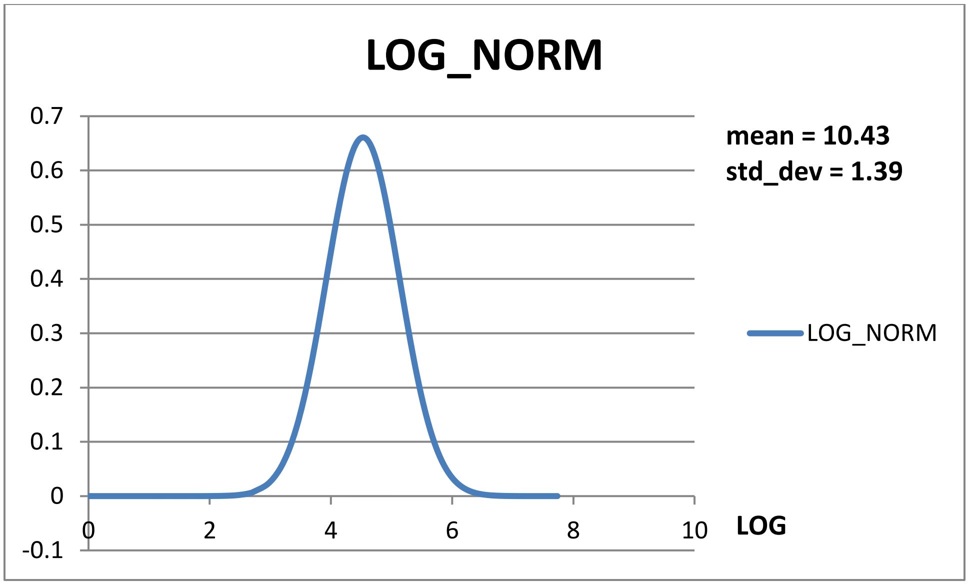


Fig#5 First Digit Test

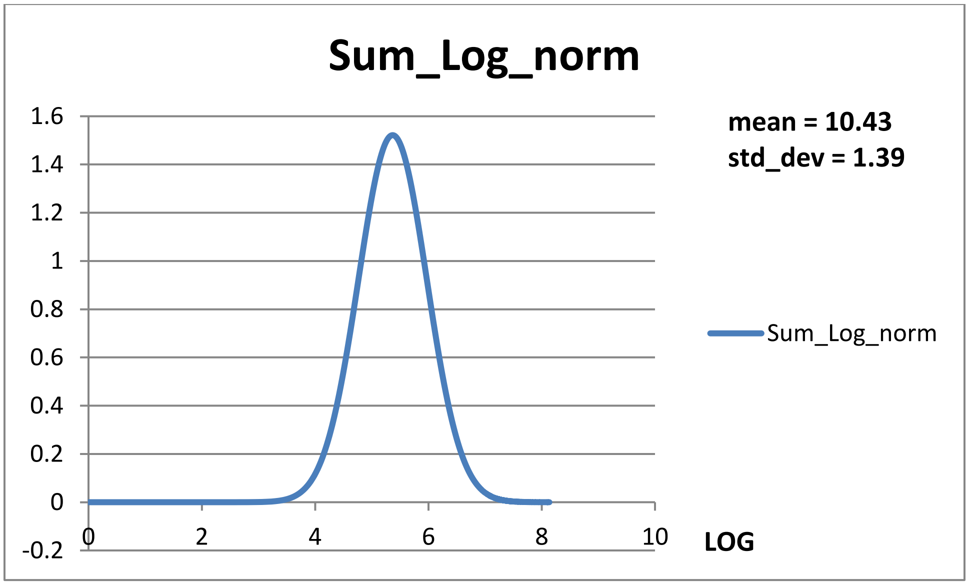
Results displayed here were derived from a Visual Basic computer program written by myself. This program was applied to the same data sources that Dr. Mark Nigrini utilized in his book, *Benford's Law*, and achieved identical results.



Fig#6 Summation Test



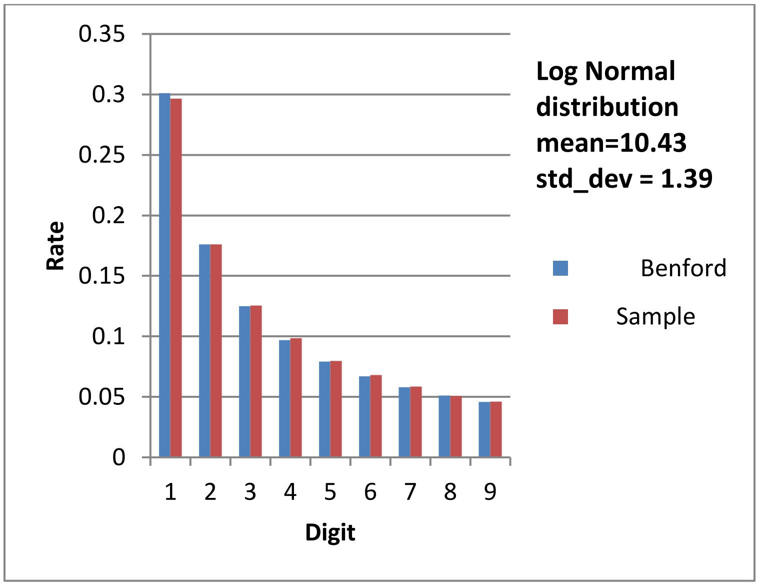
Fig#7 Plot of the probability density function of the logarithm of a data set that conforms to a Log Normal distribution



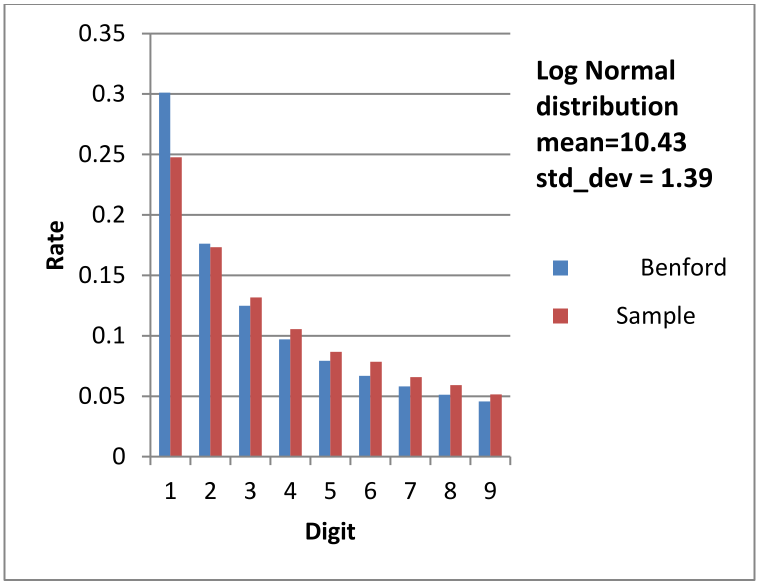
Fig#8 – Plot of the logarithm of the probability density function of the expected value (or sum) of a data set that conforms to a Log Normal distribution

	mean =		std_dev =
	10.43		1.39
Ist Digits		*1*	*0.306111816*
Summation		*2*	*0.176507538*
		3	*0.12438689*
		4	*0.096069839*
		5	*0.078274469*
		6	*0.066059837*
		7	*0.057157969*
		8	*0.05038172*
		9	*0.045049922*

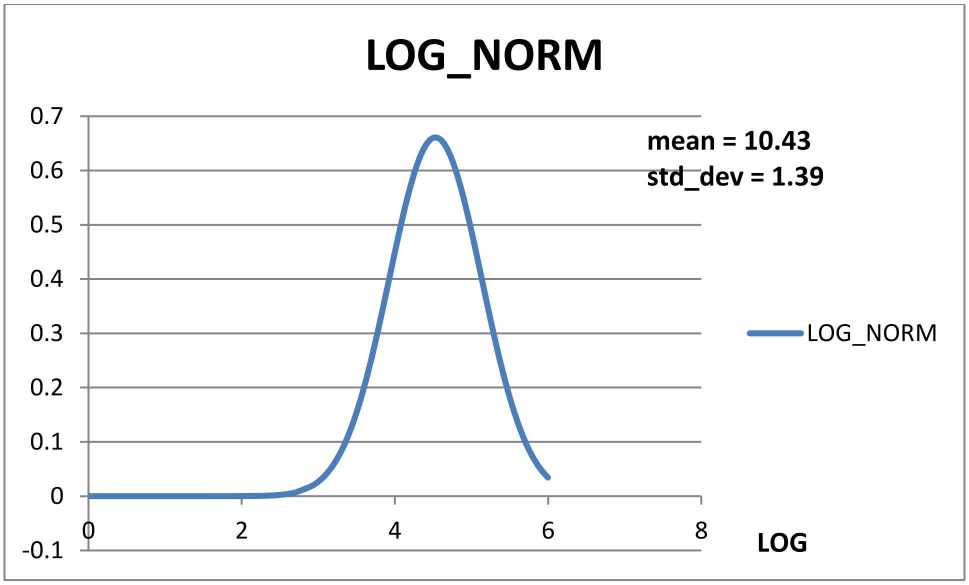
Fig#9 - Actual value derived from actual integration of the probability density function of the expected value of a log Normal distribution



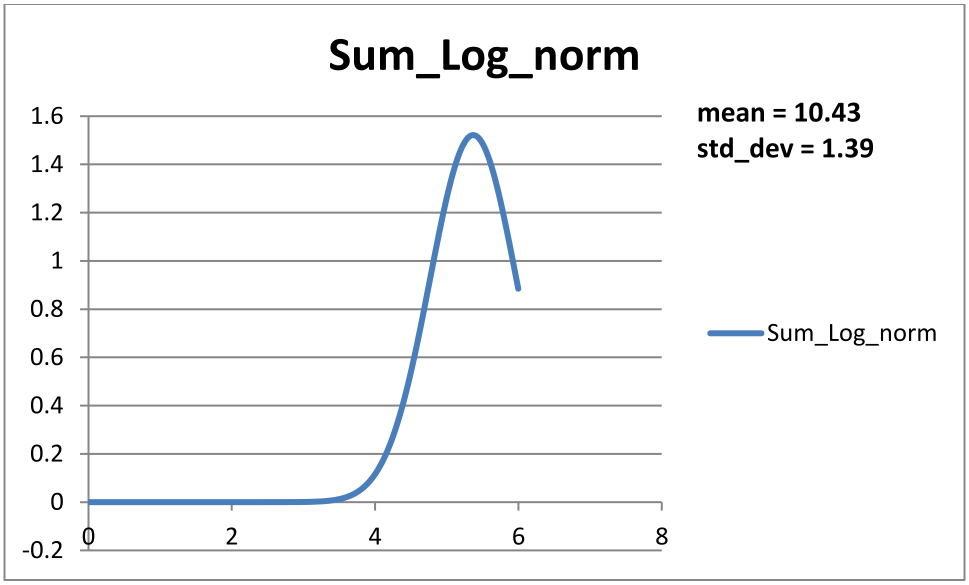
Fig#10 First Digit Test - data limit = 1,000,000



Fig#11 – Summation Test – data limit = 1,000,000



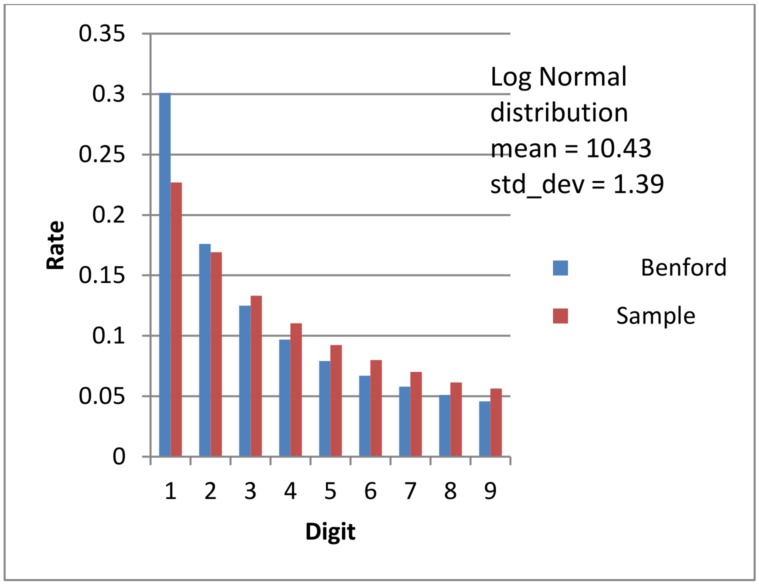
Fig#12 - Plot of the probability density function of the logarithm of a data set that conforms to a Log Normal distribution and data limited to 1,000,000



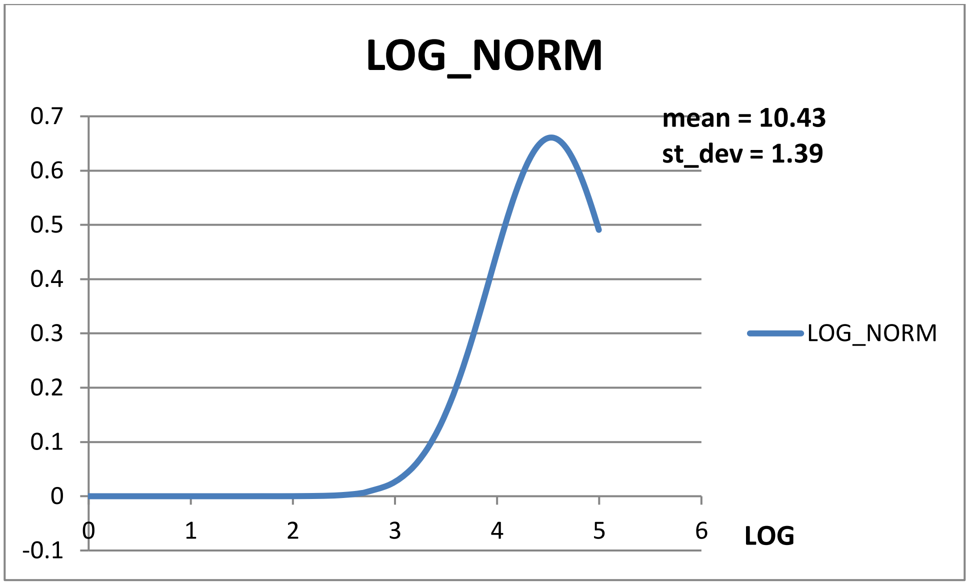
Fig#13 - – Plot of the logarithm of the probability density function of the expected value (or sum) of a data set that conforms to a Log Normal distribution and data limited to 1,000,000

Ist Digits	1	0.250422193
Summation	2	0.173952267
	3	0.131642613
	4	0.105413894
	5	0.087716646
	6	0.075027644
	7	0.065507763
	8	0.058111845
	9	0.052205135

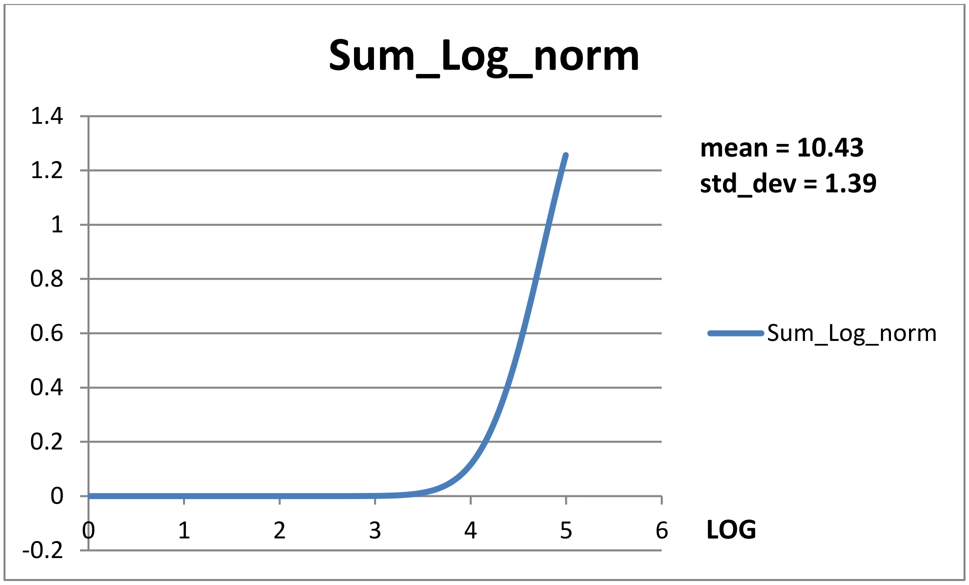
Fig#14 – Actual value derived from the actual integration of the probability density function of the expected value of the Log Normal distribution of the expected value of the Log Normal distribution



Fig#15 – First Digit Test – data limit = 100,000



Fig# 16 - Plot of the probability density function of the logarithm of a data set that conforms to a Log Normal distribution and data limited to 100,000



Fig# 18 – Plot of the logarithm of the expected value of the data set that conforms to a Log Normal distribution and data limited to 100,000

Ist Digits Summation		
	1	0.098446589
	2	0.117578991
	3	0.122161729
	4	0.12103901
	5	0.011751021
	6	0.113004286
	7	0.108183035
	8	0.103366471
	9	0.098709675

Fig#19 Actual value derived the actual integration of the probability density function of the expected value of a Log Normal Distribution

By observing the data derived from these plots it is apparent that the summation test results in a Benford distribution and not a uniform distribution. However, as the data is truncated the distribution tends to become more uniform as more data in truncated. This is because the probability density function of the logarithm of the data consists of mainly the back side (or ascending side) of the normally ascending and descending portion of the pdf curve. This is the characteristic of the logarithm of a data set that conforms to a uniform distribution (x*ln(10) * constant), which is a rising and not falling straight line.

Conclusion

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.

References:

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

Kossovski, AE (2014*) Benford's Law: Theory , the General Law of relative Quantities, and Forensic Fraud Detection Applications*, **World Scientific Publishing Company: Singapore, ISSN/ISBN 978-981-4583-68-8**

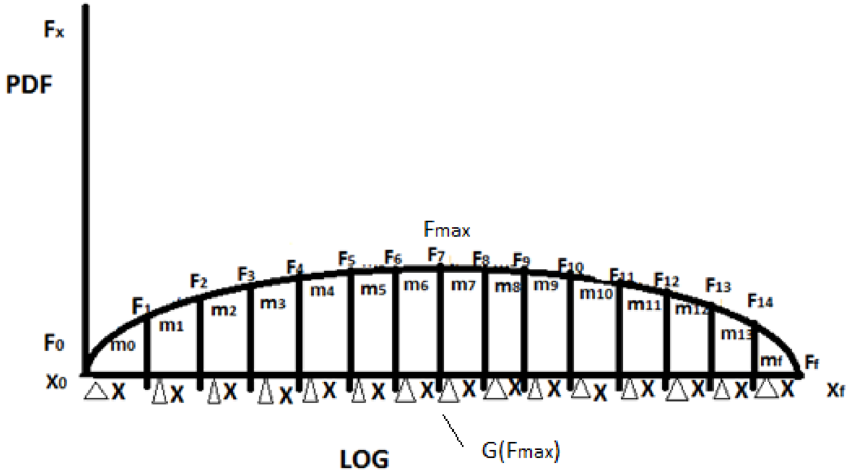
Nigrini, MJ (2012), *Benson's Law: Applications for Forensic Accounting, and Fraud Detection*, **John Wiley and Sons, ISSN/ISBN: 978-1-118-15285-0**

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

Appendix A

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. $\frac{dy}{dx}$ = $\frac{1}{xln(10)}$
   e. $\frac{dx}{dy}$ = xln(10)
   f. pdf(y) = $\frac{xln(10)}{xln(10)}$ = 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 $Y' = \frac{1}{Xf-Xo}\int_{Xo}^{Xf} Y'dx \ = \ \frac{0}{Xf-Xo}$

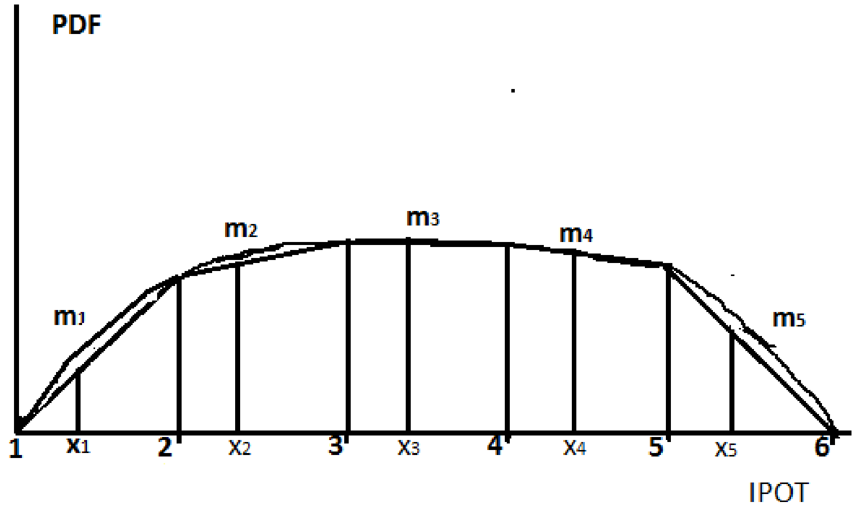
8) $F'_i\,(x) = \frac{F(i+1)-F(i)}{\Delta x}$; $\Delta x \rightarrow 0$

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

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 probability density functions of each mantissa value contained within all integral powers of ten respectively is equal to 1, which constitutes a uniform distribution

Which is the definition of a Benford distribution.

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

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

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

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

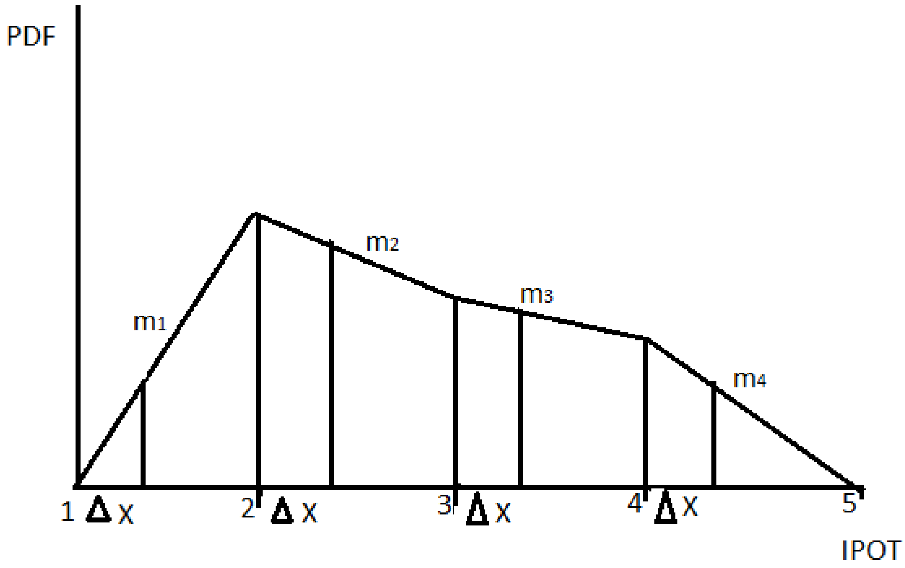
31) $(n\text{-}2)m_1 + (n\text{-}1)m_2 + ... + m_{r-2} = 1$

32) $K=1$

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

34) The resultant probability density function of the mantissas is a uniform distribution whose amplitude is equal to 1 and therefore a Benford distribution.

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 ten values approaches infinity then the sum of probability distributions of all fixed intervals from all IPOT (ΔX) equals the interval Itself (ΔX).



1) $\sum_1^4 \int_i^{i+\Delta} pdf\ dx = \frac{1}{2}\mathrm{m}_1(\Delta\mathrm{x})^2 + \mathrm{m}_1\Delta\mathrm{x}\ + \frac{1}{2}\mathrm{m}_2(\Delta\mathrm{x})^2 + (\mathrm{m}_1 + \mathrm{m}_2)\Delta\mathrm{x} + \frac{1}{2}\mathrm{m}_3(\Delta\mathrm{x})^2 + (\ \mathrm{m}_1 + \mathrm{m}_2 + \mathrm{m}_3)\Delta\mathrm{x}\ + \frac{1}{2}\mathrm{m}_4(\Delta\mathrm{x})^2 = \mathrm{K}$

2) $\frac{1}{2}(\Delta\mathrm{x})^{\ 2}\ (\mathrm{m}_1 + \mathrm{m}_2 + \mathrm{m}_3 + \mathrm{m}_4\ ) + (3m_1\ + 2\mathrm{m}_2 + \mathrm{m}_3)\Delta\mathrm{x} = \mathrm{K}$

3) $\mathrm{m}_1 + \mathrm{m}_2 + \mathrm{m}_3 + \mathrm{m}_4 = 0$

4) $3\mathrm{m}_1 + 2\mathrm{m}_2 + \mathrm{m}_3 = 1$

5) $(3m_1\ + 2\mathrm{m}_2 + \mathrm{m}_3)\Delta\mathrm{x} = \Delta\mathrm{x}$

6) $\sum_1^4 \int_i^{i+\Delta x} pdf\ dx = \Delta\mathrm{x}$

In General:

7) $\sum_{i=1}^{r-1} \int_i^{i+\Delta x} pdf\ dx\ =\ \frac{1}{2}(\Delta\mathrm{x})^2(\ \mathrm{m}_1 + \mathrm{m}_2 + \mathrm{m}_3 + \ldots + \mathrm{m}_{\mathrm{r}-1}\ ) +$

8) $[(n-2)m_1 + (n-1)m_2 + \ldots + \mathrm{m}_{\mathrm{r-2}}]\Delta\mathrm{x} \ = \Delta\mathrm{x}$

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.