Statistical Relationships Involving Benford’s Law, the Lognormal Distribution, and the Summation Theorem

R C Hall, MSEE, BSEE

e-mail: rhall20448@aol.com

Abstract

Regarding Benford’s law, many believe that the statistical data resulting from various data sources follow a Benford law probability density function ($\frac{1}{xLn(10)}$) when, in actuality, it really follows a Lognormal probability density function. The only data that strictly follows a Benford’s law probability density function is an exponential function i.e. N^x. The other sets of data conform to a Lognormal distribution and, as the standard deviation approaches infinity, approximates a true Benford distribution.

Also, the so called Summation theorem whereby the sum of the values with respect to the first digits is a uniform distribution only applies to an exponential function. The data derived from the aforementioned Lognormal distribution is more likely to conform to a Benford like distribution as the data seems to indicate.

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

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

2) Let v = Log_{10} (y) = x Log_{10} (10) = x , which is the probability distribution 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) $\text{pdf}_Y = \text{pdf}_v \times \frac{dv}{dy}$

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

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

8) $\text{pdf}_Y = \frac{1}{yln(10)}$

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

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

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

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

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

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

Let b = 6, a = 5; $\text{Log}_{10}\left(\frac{6}{5}\right) = 0.096910$

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

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

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

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

The Ist digit distribution conforms to Benford's Law.

Scale Invariance

The scale invariance associated with Benford's law states if the original data were multiplied by a constant the Ist digits distribution would still apply. 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. $\frac{1}{\ln(10)}\int_{a10^{\mathrm{N}}}^{a10^{\mathrm{N+1}}}\frac{dx}{x}$ = $\frac{1}{\ln(10)}$ [ln(a10^{N+1}) – ln(a10^{N})] = $\frac{1}{\ln(10)}$[(N+1)ln(10) + ln(a) – Nln(10) – ln(a)] = $\frac{1}{\ln(10)}$[Nln(10) - Nln(10) + ln(a) – ln(a) + ln(10)] = $\frac{\ln(10)}{ln(10)}$ = 1

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

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

Example: converting inches to centimeters

Scale factor: a = 2.54 centimeters/inch

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

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

The sum of the values with respect to the first digits for a pure Benford distribution (pdf = $\frac{1}{xln(10)}$) is a uniform distribution. ***This only applies to an exponential function i.e. N^{x}***. The following argument is proof of this assertion.

An Alternate Proof of the Summation Theorem

For numbers that follow Benford's law the sum of all numbers that start with a particular digit are the same as the sum of all numbers that start with any other digit i.e. sum of all numbers that start with 1 is the same as all numbers that start with 2 or 3, 4 etc.

The pdf of numbers begin with digit 1 is $\frac{1}{\ln(2)x}$ since $\int_1^2 \frac{dx}{x}$ = ln(2) i.e. $\frac{1}{\ln(2)}\int_1^2 \frac{dx}{x}$ = 1

Average value of x = $\int_a^b xpdf_X dx$ / $\int_a^b pdf_X dx$; **Average value of numbers between 1 and 2 =** $\frac{1}{\ln(2)}\int_1^2 \frac{xdx}{x}$ = $\frac{1}{\ln(2)}\int_1^2 dx$ = $\frac{2-1}{\ln(2)}$ = $\frac{1}{\ln(2)}$ = **1.442695**

Average value between 10 – 20 = $\frac{1}{\ln(2)}\int_{10}^{20} \frac{xdx}{x}$ = $\frac{1}{\ln(2)}\int_{10}^{20} dx$ = $\frac{20-10}{\ln(2)}$ = $\frac{10}{\ln(2)}$ = 14.42695

Generally: Average Value = $\frac{1}{\ln(2)}\int_{10^N}^{2x10^N} \mathrm{dx}$ = $\frac{10^N}{\ln(2)}$

Likewise for numbers starting with 8: pdf = $\frac{1}{xln(\frac{9}{8})}$; ***Average value =*** $\boldsymbol{\frac{1}{ln(\frac{9}{8})}\int_{8x10^N}^{9x10^N} dx = \frac{10^N}{ln(\frac{9}{8})}}$

The average value of all numbers that begin with a particular digit X the number of numbers that begin with the same digit = the sum of all numbers that begin with the same digit.

Let N = the total numbers or samples considered in a set of numbers that conform to Benford's law. If the range from 1 to 10 then the average number beginning with 1 is 1/ln(2) and the numbers of numbers that begin with 1 according to Benford's law is N X $\frac{\ln(2)}{\ln(10)}$ or N X $\log_{10} 2$. The sum of all numbers starting with 1 is $\frac{1}{\ln(2)}$ X $\frac{Nln(2)}{\ln(10)}$ = $\boldsymbol{\frac{N}{ln(10)}}$.

The situation is a little different for numbers spread over several orders of magnitude.

Consider a range from 1 to 100,000. $\int_1^{100,000} \frac{dx}{x}$ = ln(10^5) = 5ln(10); pdf = $\frac{1}{5\ln(10)}$

Numbers between 1-2 = $\frac{N\int_1^2 \frac{dx}{x}}{5\ln(10)}$ = $\frac{Nln(2)}{5\ln(10)}$ = $\frac{N}{5}\log_{10} 2$

Numbers between 10 – 20 = $\frac{N\int_{10}^{20} \frac{dx}{x}}{5\ln(10)}$ = $\frac{N}{5}\log_{10} 2$

Numbers between 100 – 200 = $\frac{N\int_{100}^{200} \frac{dx}{x}}{5\ln(10)}$ = $\frac{N}{5}\log_{10} 2$

Numbers between 1,000 – 2,000 = $\frac{N\int_{1000}^{2000}\frac{dx}{x}}{5\ln(10)}$ = $\frac{N}{5}\log_{10}2$

Numbers between 10,000 – 20,000 = $\frac{N\int_{10000}^{20000}\frac{dx}{x}}{5\ln(10)}$ = $\frac{N}{5}\log_{10}2$

Total = $\frac{N\log_{10}2}{5}$ X 5 = N$\log_{10}2$

The average value between 1 – 2 = $\frac{1}{\ln(2)}$ = 1.44269504

The average value between 10 – 20 = $\frac{10}{\ln(2)}$ = 14.4269504

The average value between 100 – 200 = $\frac{100}{\ln(2)}$ = 144.4269504

The average value between 1,000 – 2,000 = $\frac{1000}{\ln(2)} =$ 1444.269504

The average value between 10,000 – 20,000 = $\frac{10000}{\ln(2)}$ = 14442.69504

Summation = average value X the number of samples starting with the number 1

1 – 2: ($\frac{1}{\ln(2)}$)X($\frac{\mathrm{N}\ln(2)}{5\ln(10)}$) = $\frac{N}{5\ln(10)}$

10 – 20: ($\frac{10}{\ln(2)}$)X($\frac{Nln(2)}{5\ln(10)}$) = $\frac{10N}{5\ln(10)}$

100 – 200: ($\frac{100}{\ln(2)}$)X($\frac{Nln(2)}{5\ln(10)}$) = $\frac{100N}{5\ln(10)}$

1,000 – 2,000: ($\frac{1000}{\ln(2)}$)X($\frac{Nln(2)}{5\ln(10)}$ = $\frac{1000N}{5\ln(10)}$

10,000 – 20,000: ($\frac{10000}{\ln(2)}$)X($\frac{Nln(2)}{5\ln(10)}$) = $\frac{10000N}{5\ln(10)}$

Total Summation = $\frac{N}{5\ln(10)}$ X (1 +10 +100 + 1000 + 10000) = $\frac{N}{5\ln(10)}$ X 11,111

In General: Summation (assuming (highest value/lowest value) mod 10 = 0) = $\frac{N}{\log_{10}\frac{\max value}{\min value}}$ X($\frac{1}{\ln(10)}$) X $\sum_{k=\log_{10}\min value}^{\log_{10}\max value-1}10^{\mathrm{k}}$

For numbers that range between 1.0 – 1.1, 10 – 11, 100 – 110 etc, the average values change as follows:

The average value between 1.0 – 1.1 = $\frac{1}{ln1.1}\int_{1.0}^{1.1}\frac{xdx}{x}$ = $\frac{1.1-1.0}{ln1.1}$ = $\frac{0.1}{ln1.1}$ = 1.0492

Likewise for:

10 – 11: $\frac{1}{ln1.1}$ = 10.4921

100 - 110: $\frac{10}{ln1.1}$ = 104.9205

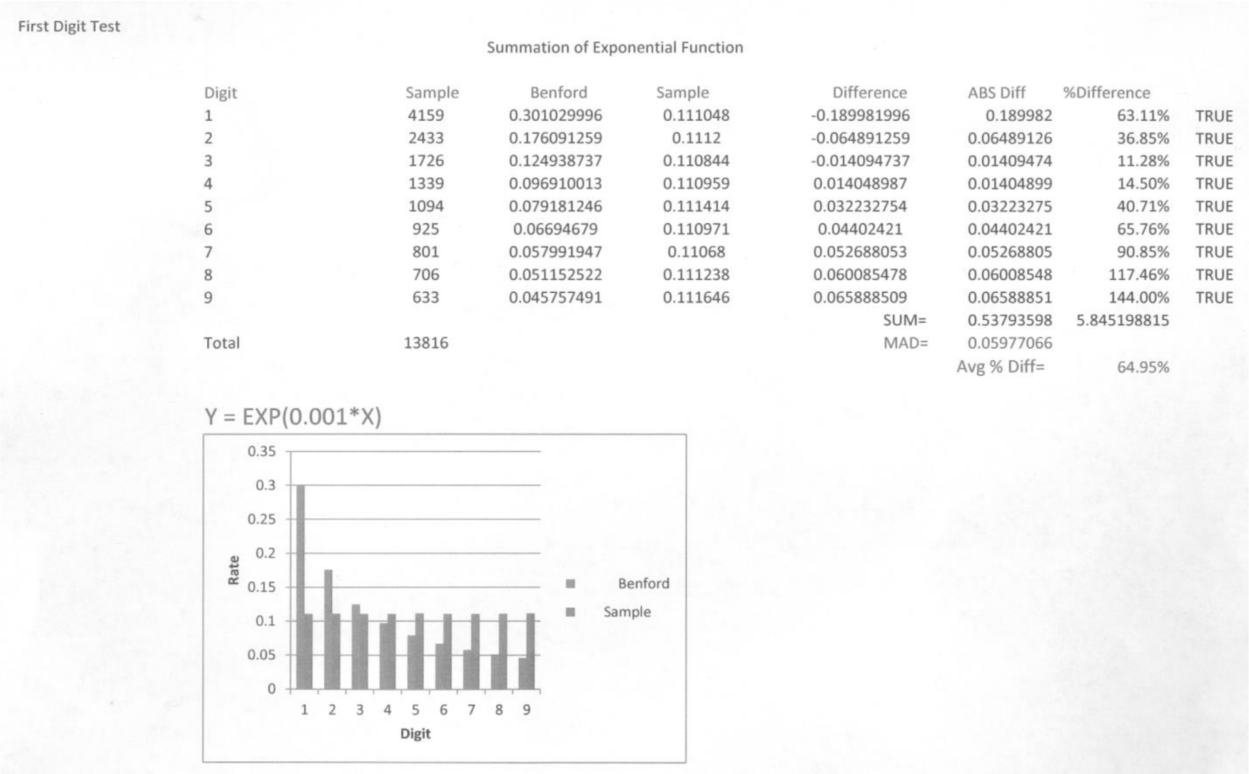
1,000 – 1100: $\frac{100}{ln1.1}$ = 1,049.2059

10,000 – 20,000: $\frac{1000}{ln1.1}$ = 10,492.059

Etc.

In general: Summation = $\frac{N}{\log_{10}\frac{\max value}{\min value}}$ X $\left(\frac{1}{\ln(10)}\right)$ X 0.1$\sum_{k=\log_{10}\min value}^{\log_{10} max\ value-1} 10^{k}$

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



First Digit Test

Summation of Exponential Function

Digit	Sample	Benford	Sample	Difference	ABS Diff	%Difference	
1	4159	0.301029996	0.111048	-0.189981996	0.189982	63.11%	TRUE
2	2433	0.176091259	0.1112	-0.064891259	0.06489126	36.85%	TRUE
3	1726	0.124938737	0.110844	-0.014094737	0.01409474	11.28%	TRUE
4	1339	0.096910013	0.110959	0.014048987	0.01404899	14.50%	TRUE
5	1094	0.079181246	0.111414	0.032232754	0.03223275	40.71%	TRUE
6	925	0.06694679	0.110971	0.04402421	0.04402421	65.76%	TRUE
7	801	0.057991947	0.11068	0.052688053	0.05268805	90.85%	TRUE
8	706	0.051152522	0.111238	0.060085478	0.06008548	117.46%	TRUE
9	633	0.045757491	0.111646	0.065888509	0.06588851	144.00%	TRUE
				SUM=	0.53793598	5.845198815	
Total	13816			MAD=	0.05977066		
				Avg % Diff=		64.95%	

The distribution with respect to the Ist digits is a uniform distribution

Proof that the multiplication of statistically independent numbers results in a Lognormal distribution and the resulting distribution approaches a Benford distribution as the standard deviation approaches infinity.

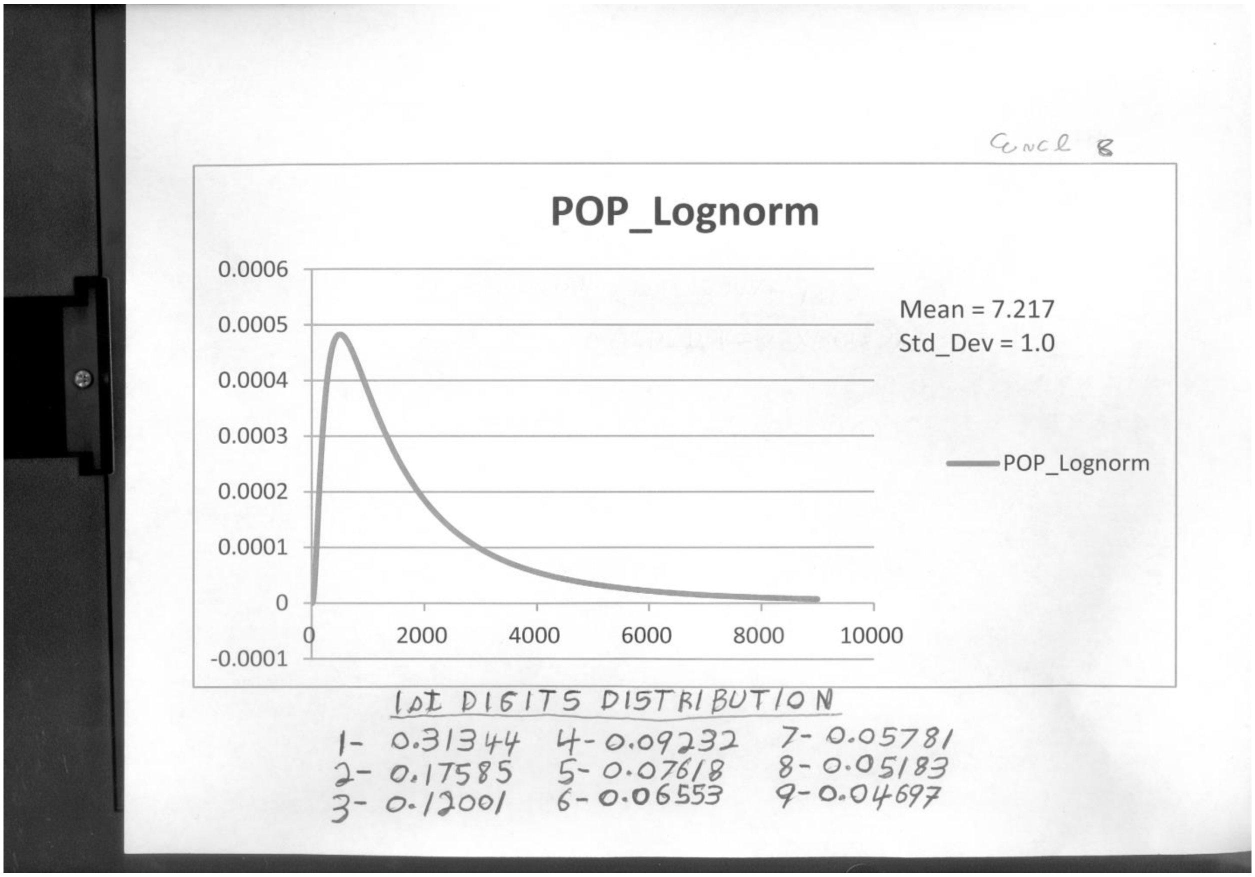
Most numbers encountered in real life such as populations, scientific data, and accounting data are derived from the multiplication the multiplication of statistically independent numbers, which constitute a Lognormal (as opposed to a normal distribution) analogous to a Gaussian or Normal probability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

The following is a proof that the multiplication of statistically independent numbers result in a Lognormal distribution and as its standard deviation approaches infinity the probability density function approaches a Benford distribution.

1) Y = the product of x_1, x_2 x_3, x_4 ……x_r
2) Ln(Y) = Ln(x_1) + Ln(x_2) + Ln(x_3) + Ln(x_4) + …. Ln(x_r)
3) Because Ln(x) is a function of x and Lx(x_i) are statistically independent
4) Pdf conforms to the Central Limit Theorem as r → ∞.
5) pdf$_Y$ = $\frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-(y-u)^2/2\sigma^2}$
6) Y = Ln(x); dy = $\frac{dx}{x}$
7) pdf$_Y$ dy = pdf$_X$dx
8) pdf$_X$= pdf$_Y\frac{dy}{dx}$ = $\frac{pdf_Y}{x}$
9) pdf$_X$= $=\quad \frac{1}{x\sqrt{2\pi\sigma^2}}\ e^{-(y-u)^2/2\sigma^2}\qquad \frac{1}{x\sqrt{2\pi\sigma^2}}\ e^{-(Ln(x)-u)^2/2\sigma^2}$; σ is the standard deviation of Ln(x)
10) The Benford probability density function = $\frac{1}{xLn(10)}$.
11) The Lognormal probability density function = $\frac{1}{x\sqrt{2\pi\sigma^2}}\ e^{-(Ln(x)-u)^2/2\sigma^2}$
12) Let u = 0
13) For x=1: 1/x = 1; $\frac{1}{x\sqrt{2\pi\sigma^2}}\ e^{-(Ln(x))^2/2\sigma^2}$ =$\frac{1}{\sqrt{2\pi\sigma^2}}$
14) Normalize by multiplying the Lognormal distribution by $\frac{\sqrt{2\pi\sigma^2}}{Ln(10)}$
15) = $\frac{1}{xLn(10)}\ e^{-(Ln(x)-u)^2/2\sigma^2}$
16) For any given value of x the value $e^{-(Ln(x)-u)^2/2\sigma^2}$ approaches 1 as σ approaches ∞

Figs# 2-4 Illustrate the shape of the Lognormal probability density function as the standard deviation increases and eventually approaches a Benford probability density function

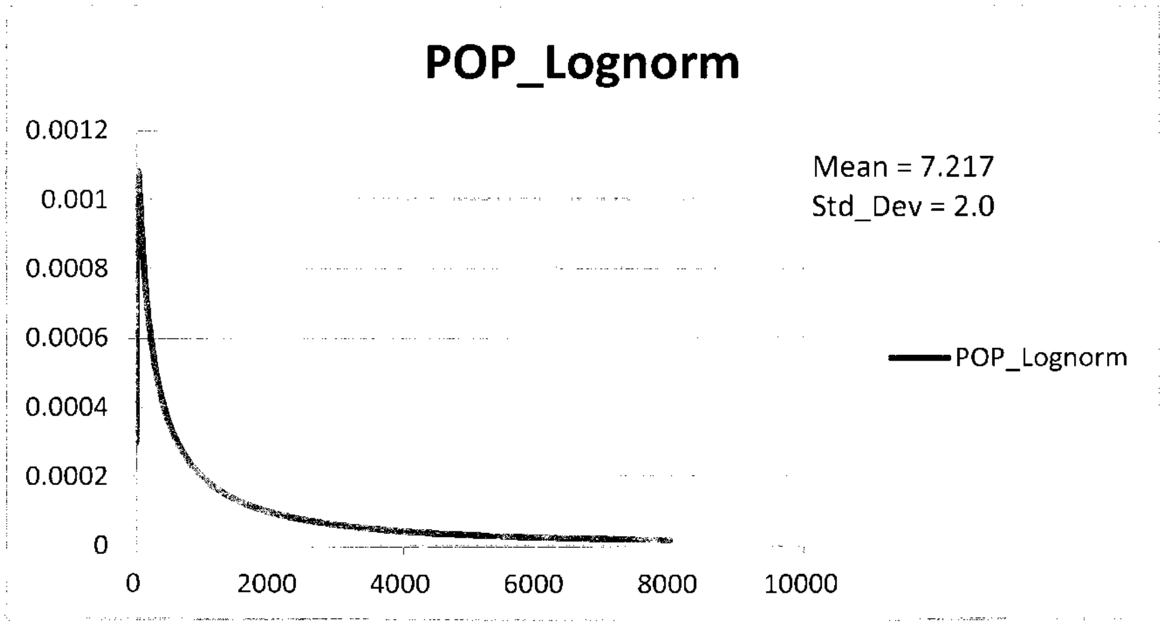
Fig#2 – Lognormal Probability Density Function v. Benford Probability Density Function



1ST DIGITS DISTRIBUTION

1- 0.31344	4- 0.09232	7- 0.05781
2- 0.17585	5- 0.07618	8- 0.05183
3- 0.12001	6- 0.06553	9- 0.04697

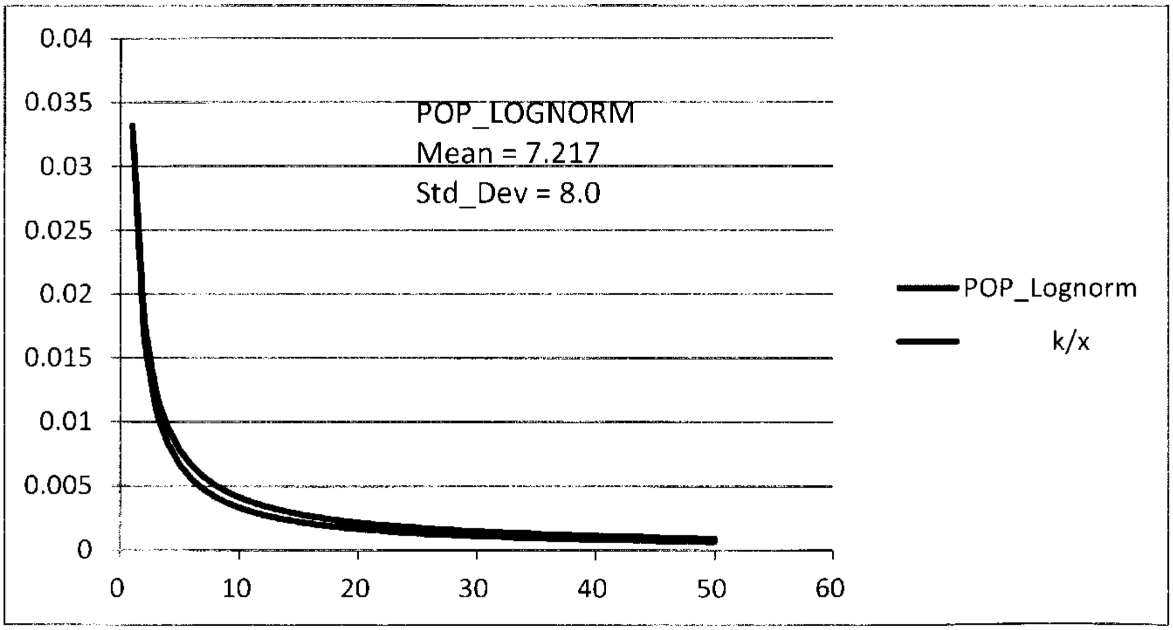
Fig#3 – Lognormal Probability Density Function v. Benford Probability Density Function



1ST DIGITS DISTRIBUTION

1- 0.30103	4- 0.09691	7- 0.05799
2- 0.17609	5- 0.07918	8- 0.05115
3- 0.12494	6- 0.06695	9- 0.04576

Fig#4 – Lognormal Probability Density Function v. Benford Probability Density Function



1st DIGITS DISTRIBUTION

1- 0.30103 4- 0.09691 7- 0.05799
2- 0.17609 5- 0.07918 8- 0.05115
3- 0.12494 6- 0.06695 9- 0.04576

Also, if the standard deviation of the Lognormal probability density function approaches 0 then the distribution approaches a Gaussian or Normal distribution.

Proof that as the standard deviation of a lognormal distribution approaches 0 the distribution becomes a Normal distribution with a mean of e^{u} where u is the mean of the natural logarithm of the data.

1) Lognormal distribution: F(x) = $\frac{1}{x\sqrt{2\pi\sigma^2}} e^{\frac{-(Ln(x)-u)^2}{2\sigma^2}}$; u = mean(ln(x)), σ = std_dev(ln(x))

2) Determine the mode of the Lognormal distribution i.e.

$\frac{dy}{dx} = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{dy}{dx} (\frac{e^{-(Ln(x)-u)^2/2\sigma^2}}{x}) = 0$; solve for x

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

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

5) Ln(x) = u-σ²

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

7) As σ→0; x→e^u

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

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

10) Ln(e^u) + $\frac{x- e^u}{e^u}$ - $\frac{(x-e^u)^2}{2e^{2u}}$ + $\frac{(x-e^u)^3}{3e^{3u}}$ +….+

11) Ln(x-e^u) ~ Ln(e^u) + $\frac{x- e^u}{e^u}$ as σ → 0

12) Ln(x-e^u) ~ u + $\frac{x- e^u}{e^u}$

13) F(x) ~ $\frac{1}{x\sqrt{2\pi\sigma^2}} e^{\frac{-(u+\frac{x-e^u}{e^u}-u)^2}{2\sigma^2}}$

14) F(x) = ~ $\frac{1}{e^u\sqrt{2\pi\sigma^2}} e^{\frac{-(\frac{x-e^u}{e^u})^2}{2\sigma^2}}$ as σ → 0

15) F(x) ~ $\frac{1}{\sqrt{2\pi(\sigma e^{\mathrm{u}})^2}} e^{\frac{-(x-e^{\mathrm{u}})^2}{2(\sigma e^{\mathrm{u}})^2}}$

16) **u$_x$** = mean(x); **σ$_x$** = std_dev(x)

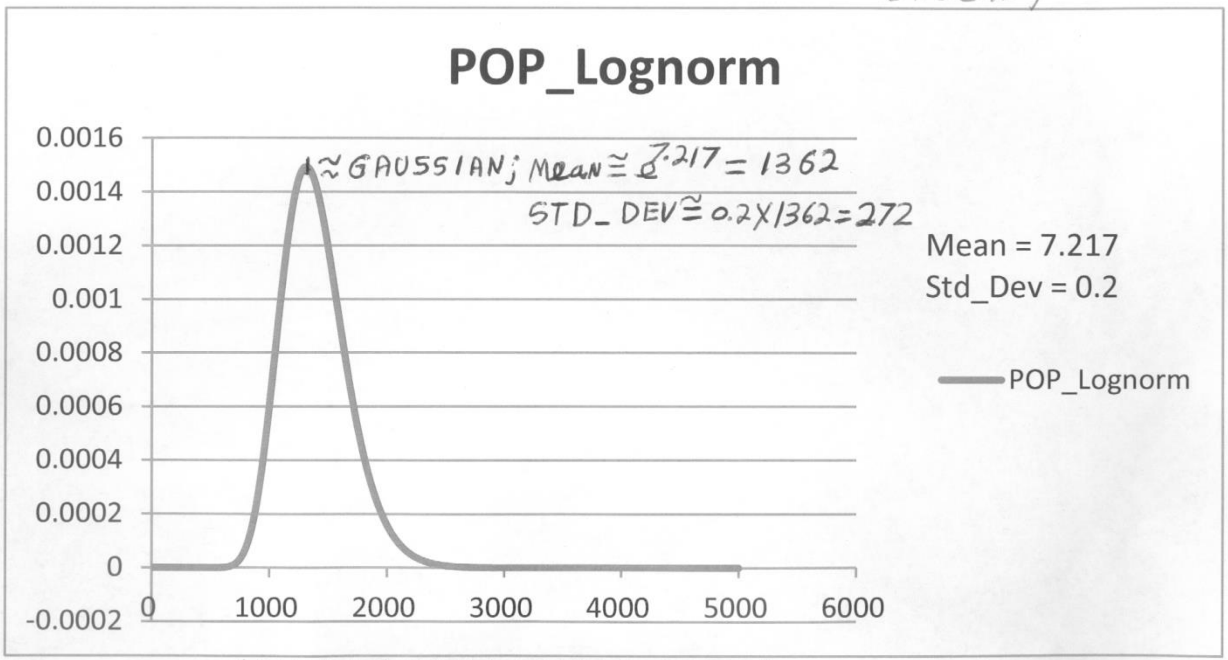
17) **u$_x$** ~ e^u ; **σ$_x$** ~ u$_x$ σ

18) F(x) ~ $\frac{1}{\sqrt{2\pi(\sigma_x)^2}} e^{\frac{-(x-u_x)^2}{2(\sigma_x)^2}}$

19) Which is a Normal Distribution with a mean of e^u

Fig#5 – Probability Density Function of a Lognormal Distribution with a Small Value Standard Deviation

Encl#9



1ST DIGITS DISTRIBUTION

1-0.91151 4-3.0508E-07 7-0.00345
2-0.02742 5-2.0357E-05 8- 0.01520
3-0. 3.9559E-05 6- 0.000414 9- 0.04194

The probability density function of an exponential function i.e. 10^x is $\frac{\ln(x)}{\ln(10)}$ while the probability function of the Log_{10} of the exponential function is a constant namely, 1.

The probability density function of logarithm of data that conforms to a Lognormal distribution is a
Gaussian or Normal distribution.

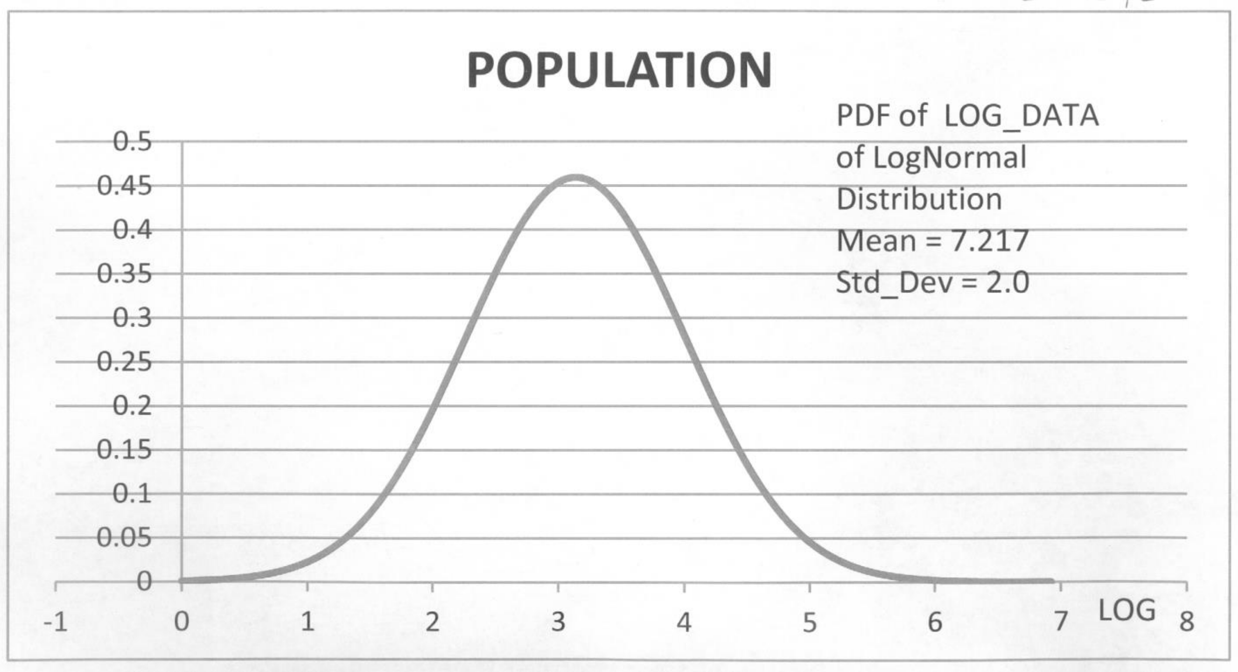
The following argument constitutes a proof that the probability density function of the logarithm of data that conforms to a Lognormal distribution is a Gaussian or Normal distribution.

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

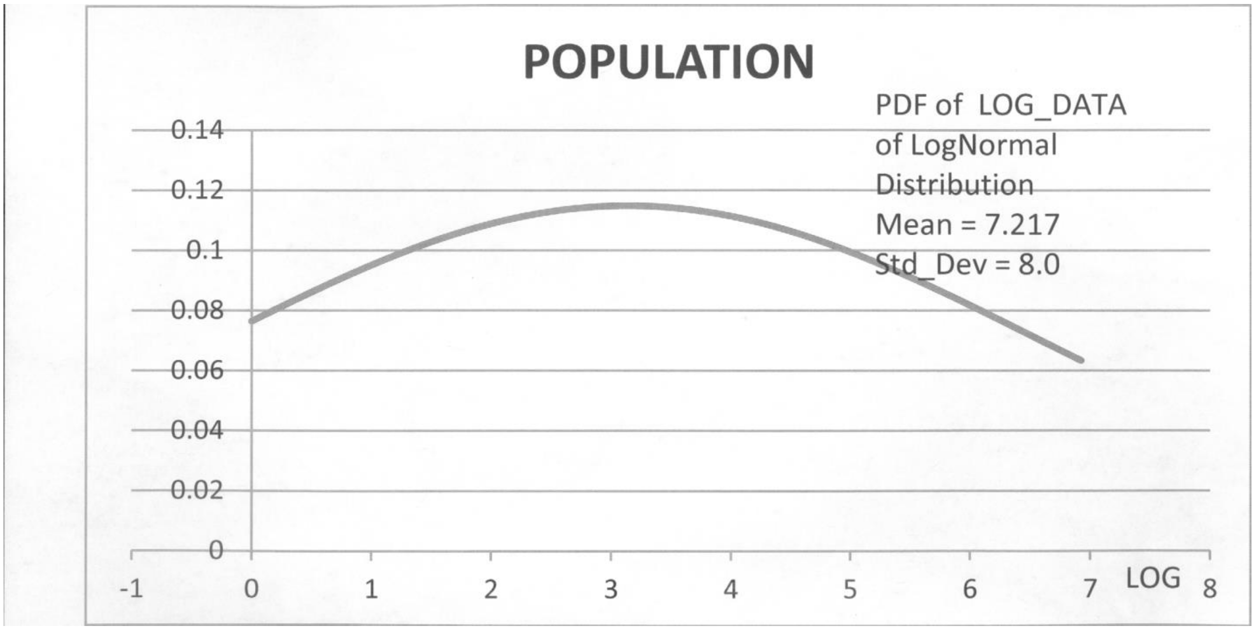
Fig#6 – Probability Density Function of the Logarithm of a Data Set that Conforms to a Lognormal Distribution

Encl 10

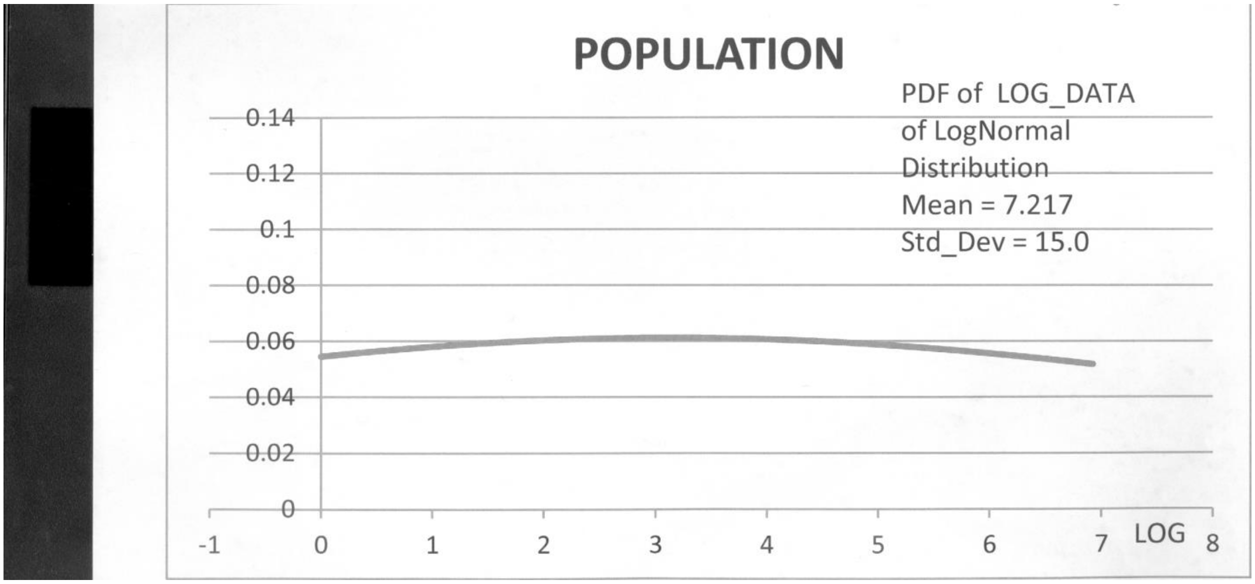


For an exponential distribution, the mantissas between integral powers of ten (IPOT) are uniform since the probability density function is 1. This accounts for the fact that numbers beginning with 1 occur about 30% of the time and numbers beginning with 9 occur about 4.6% of the time.

Fig#7 – Probability Density Function of a Data Set that Conforms to a Lognormal Distribution as the Standard Deviation Increases



Fig#8 - Probability Density Function of a Data Set that Conforms to a Lognormal Distribution as the Standard Deviation Increases



For a Lognormal distribution or any other distribution if it can be shown that the sum of all mantissas for each IPOT approaches a constant value as the number of number of integral powers of ten (IPOT) approaches infinity and therefore 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) = $\frac{\ln(x)}{\ln(10)}$

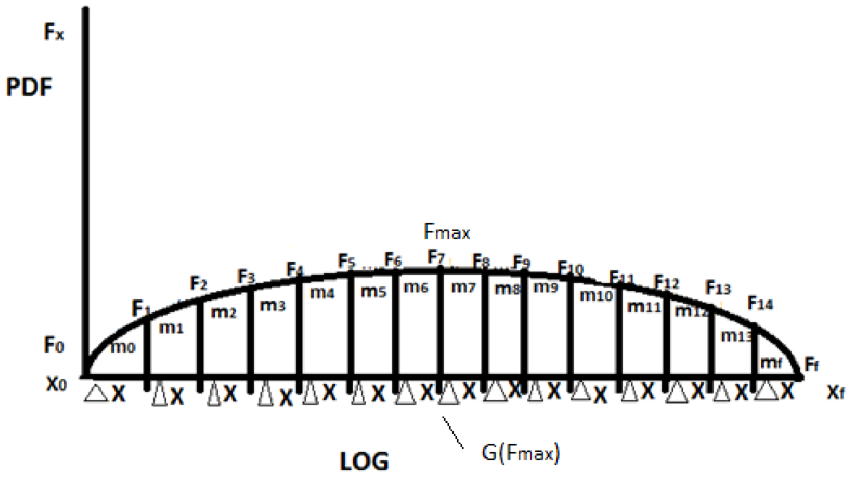
c. pdf(y) = pdf(x) $\frac{dx}{dy}$

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

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' = $\frac{1}{Xf-Xo}\int_{Xo}^{Xf} Y'dx$ = $\frac{0}{Xf-Xo}$

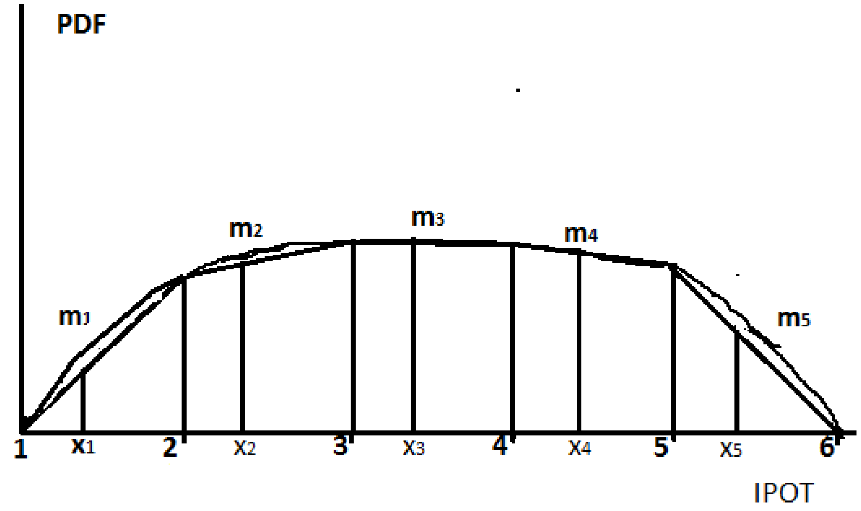
8) $F'_i\,(x) = \frac{F(i+1)-F(i)}{\Delta x}$; Δx → 0

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

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

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

Let's consider a simpler case.



12) Let ΔX = 1

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

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

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

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) $\sum_{i=1}^{5} x_{\mathrm{i}} = 4m_1 + 3m_2 + 2m_3 + m_4 = K$ (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) $\sum_{i=1}^{r-1} m_{\mathrm{i}} =$

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

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

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

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

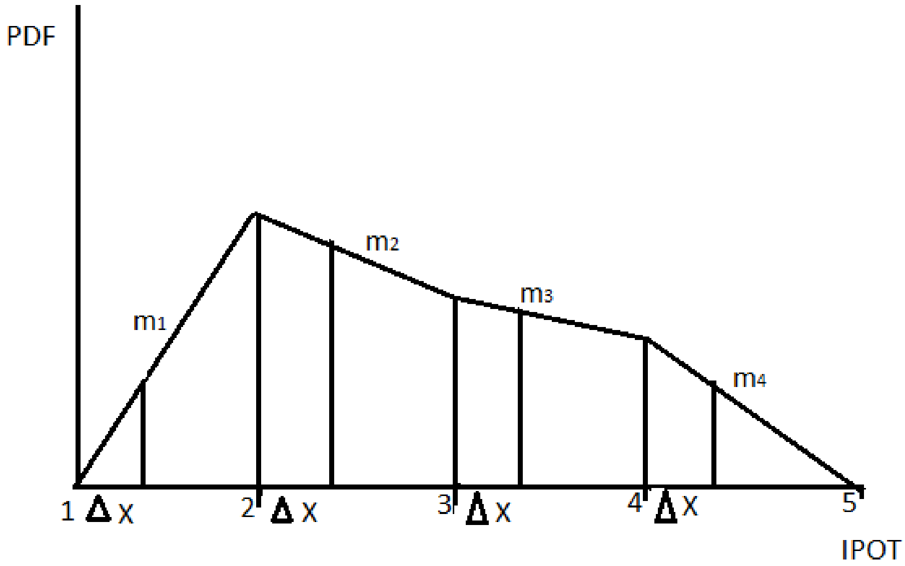
32) K=1

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 $\frac{\Delta x}{N} \rightarrow 0$.

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 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) $\frac{1}{2}\mathrm{m}_1 + \mathrm{m}_1 + + \frac{1}{2}\mathrm{m}_2 + \mathrm{m}_1 + \mathrm{m}_2 + + \frac{1}{2}\mathrm{m}_3 + \mathrm{m}_1 + \mathrm{m}_2 + \mathrm{m}_3 + \frac{1}{2}\mathrm{m}_4 =$
5) $\frac{1}{2}(\mathrm{m}_1 + \mathrm{m}_2 + \mathrm{m}_3 + \mathrm{m}_4) + 3\ \mathrm{m}_1 + 2\ \mathrm{m}_2 + \mathrm{m}_3 = 1$
6) $3\mathrm{m}_1 + 2\mathrm{m}_2 + \mathrm{m}_3 = 1$
7) $(3m_1\ + 2\mathrm{m}_2 + \mathrm{m}_3)\Delta\mathrm{x} = \Delta\mathrm{x}$
8) $\sum_1^4 \int_i^{i+\Delta x} pdf\ dx = \Delta\mathrm{x}$

In General:

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

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

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 intervali.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.

The probability distribution function of the sum of the values with respect to the first digits is a uniform distribution for an exponential function i.e. 10^x **but not for a Lognormal distribution.** The distribution is more apt to be a Benford Distribution as the following argument asserts.

Proof that the probability distribution of the sum of the values of a Lognormal 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 $\mathrm{Pdf}_X = \frac{e^{-(\ln(x)-u)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$
10. y = Log(x)
11. Pdf_Y dy = Pdf_X dx
12. $\mathrm{Pdf}_Y = \mathrm{Pdf}_X \frac{dx}{dy}$

13. $\frac{dy}{dx} = \frac{1}{xln(10)}$; $\frac{dx}{dy}$ = x ln(10)

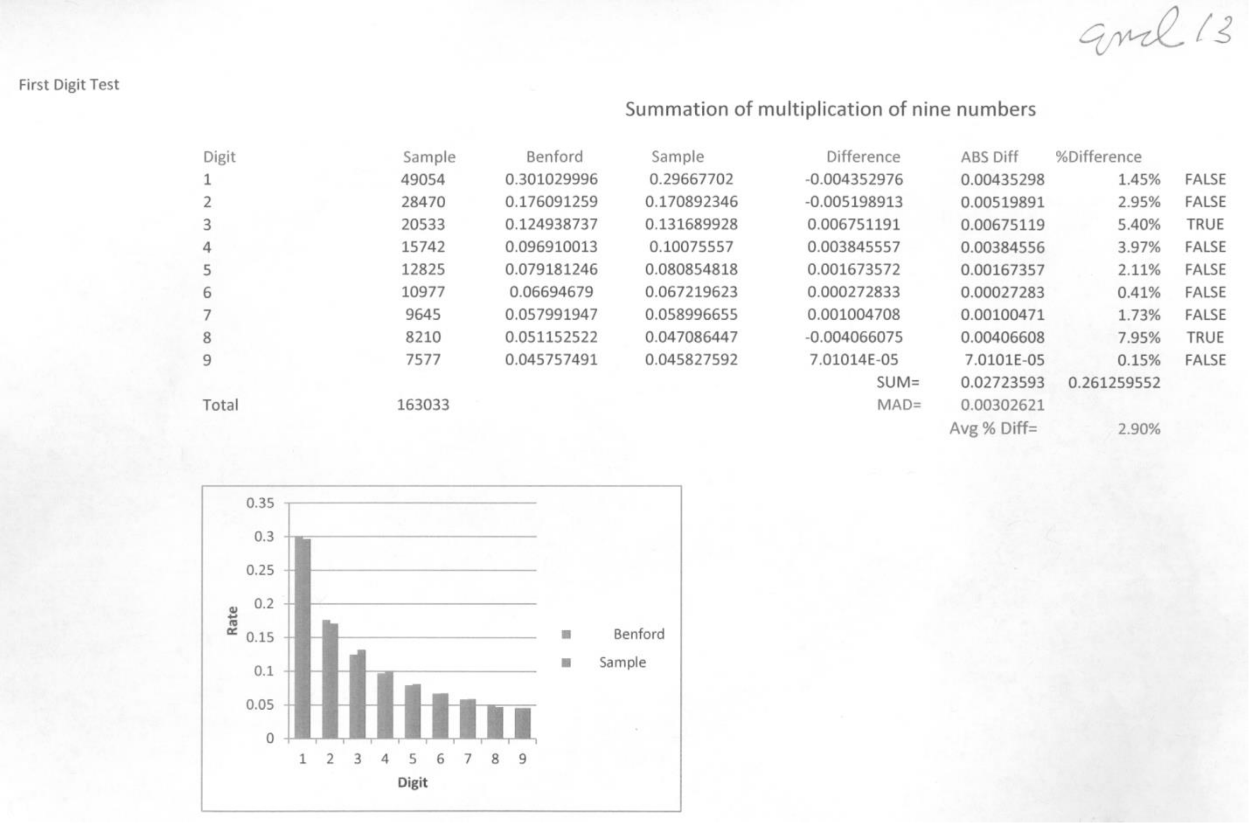
14. Pdf$_\gamma$ (log(x)) = $\ln(10)10^{\log(x)}\ \frac{e^{-\left(\ln(10^{\log(x)})-u\right)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$ =

15. (x)* ln(10)* $\frac{e^{-(\ln(x)-u)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$

16. If Log plot of can be 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.

Fig#9 – Summation with Respect to the Ist Digits i.e. 1,2,3,4,5,6,7,8,9 of the Multiplication of Nine Random Numbers



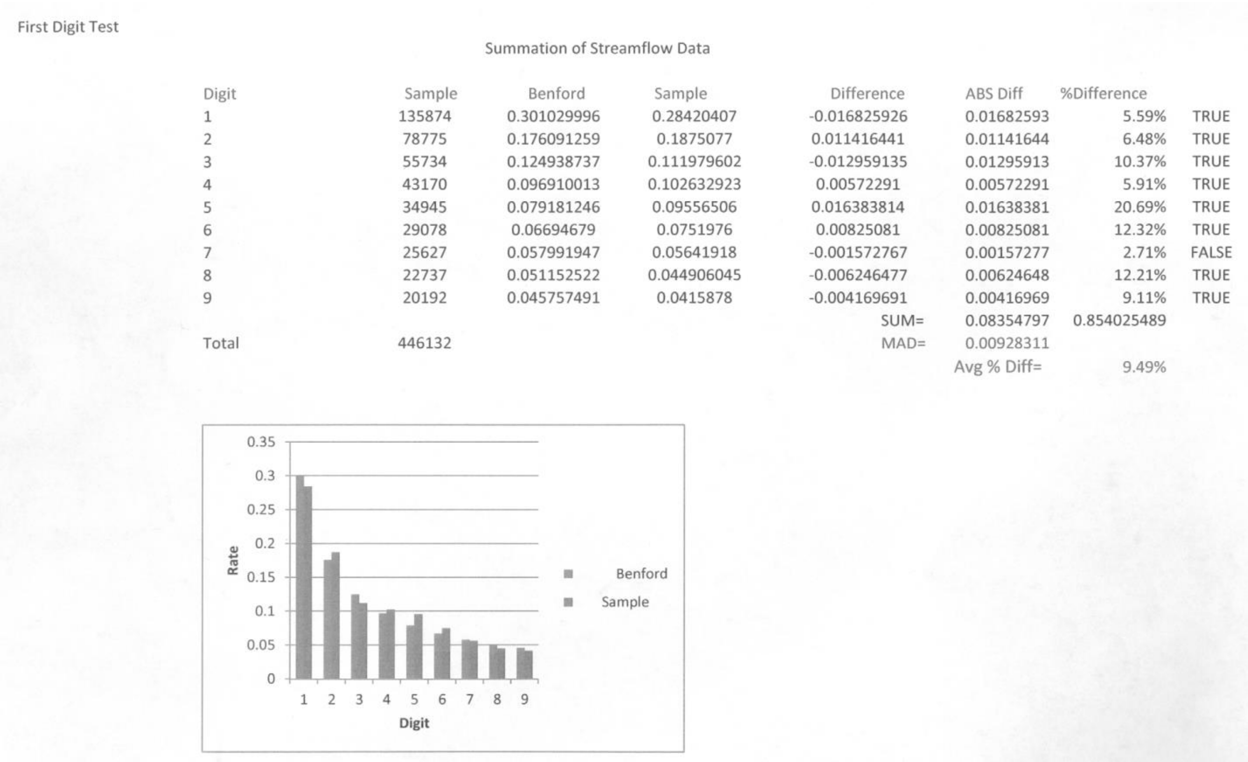
gnd 13

First Digit Test

Summation of multiplication of nine numbers

Digit	Sample	Benford	Sample	Difference	ABS Diff	%Difference	
1	49054	0.301029996	0.29667702	-0.004352976	0.00435298	1.45%	FALSE
2	28470	0.176091259	0.170892346	-0.005198913	0.00519891	2.95%	FALSE
3	20533	0.124938737	0.131689928	0.006751191	0.00675119	5.40%	TRUE
4	15742	0.096910013	0.10075557	0.003845557	0.00384556	3.97%	FALSE
5	12825	0.079181246	0.080854818	0.001673572	0.00167357	2.11%	FALSE
6	10977	0.06694679	0.067219623	0.000272833	0.00027283	0.41%	FALSE
7	9645	0.057991947	0.058996655	0.001004708	0.00100471	1.73%	FALSE
8	8210	0.051152522	0.047086447	-0.004066075	0.00406608	7.95%	TRUE
9	7577	0.045757491	0.045827592	7.01014E-05	7.0101E-05	0.15%	FALSE
				SUM=	0.02723593	0.261259552	
Total	163033			MAD=	0.00302621		
					Avg % Diff=	2.90%	

Fig#10 - Summation with Respect to the Ist Digits i.e. 1,2,3,4,5,6,7,8,9 of Stream Flow data

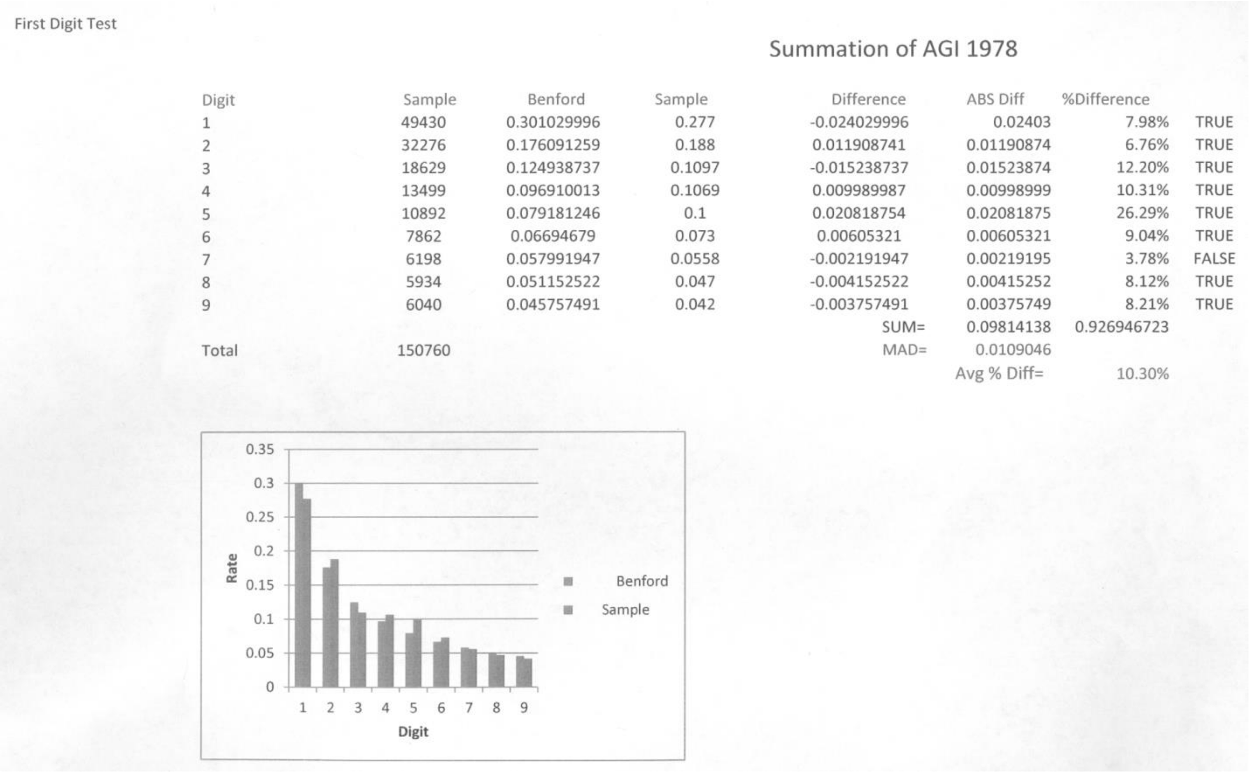


First Digit Test

Summation of Streamflow Data

Digit	Sample	Benford	Sample	Difference	ABS Diff	%Difference	
1	135874	0.301029996	0.28420407	-0.016825926	0.01682593	5.59%	TRUE
2	78775	0.176091259	0.1875077	0.011416441	0.01141644	6.48%	TRUE
3	55734	0.124938737	0.111979602	-0.012959135	0.01295913	10.37%	TRUE
4	43170	0.096910013	0.102632923	0.00572291	0.00572291	5.91%	TRUE
5	34945	0.079181246	0.09556506	0.016383814	0.01638381	20.69%	TRUE
6	29078	0.06694679	0.0751976	0.00825081	0.00825081	12.32%	TRUE
7	25627	0.057991947	0.05641918	-0.001572767	0.00157277	2.71%	FALSE
8	22737	0.051152522	0.044906045	-0.006246477	0.00624648	12.21%	TRUE
9	20192	0.045757491	0.0415878	-0.004169691	0.00416969	9.11%	TRUE
				SUM=	0.08354797	0.854025489	
Total	446132			MAD=	0.00928311		
					Avg % Diff=	9.49%	

Fig#11 - Summation with Respect to the Ist Digits i.e. 1,2,3,4,5,6,7,8,9 of IRS Collection of Adjusted Gross Income (AGI) in 1978



First Digit Test

Summation of AGI 1978

Digit	Sample	Benford	Sample	Difference	ABS Diff	%Difference	
1	49430	0.301029996	0.277	-0.024029996	0.02403	7.98%	TRUE
2	32276	0.176091259	0.188	0.011908741	0.01190874	6.76%	TRUE
3	18629	0.124938737	0.1097	-0.015238737	0.01523874	12.20%	TRUE
4	13499	0.096910013	0.1069	0.009989987	0.00998999	10.31%	TRUE
5	10892	0.079181246	0.1	0.020818754	0.02081875	26.29%	TRUE
6	7862	0.06694679	0.073	0.00605321	0.00605321	9.04%	TRUE
7	6198	0.057991947	0.0558	-0.002191947	0.00219195	3.78%	FALSE
8	5934	0.051152522	0.047	-0.004152522	0.00415252	8.12%	TRUE
9	6040	0.045757491	0.042	-0.003757491	0.00375749	8.21%	TRUE
				SUM=	0.09814138	0.926946723	
Total	150760			MAD=	0.0109046		
					Avg % Diff=	10.30%	

The distributions appear to reasonably close to a Benford distribution and **not** a uniform distribution

Conclusion

A true Benford distribution only occurs with an exponential function. All other numbers consisting of an aggregation of multiplied statistically independent numbers conform to a Lognormal distribution which approaches a true Benford distribution as the standard deviation approaches infinity.

The probability density function of the logarithm of a data set that conforms to a true Benford distribution i.e. exponential function is a constant value,

whereas the probability density function of the logarithm of a Lognormal data set is a Gaussian or Normal distribution that approaches a constant value as the standard deviation approaches infinity.

The summation with respect to the Ist digits is a uniform distribution only for exponential functions and a Benford like distribution for Log Normal distributions.

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