Comparison of the Theoretical and Empirical Results for the Benford's Law Summation Test Performed on Data that Conforms to a Log Normal Distribution.

R C Hall, MSEE, BSEE

e-mail: rhall20448@aol.com

Abstract

The Benford's Law 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 a Benford's Law distribution.

Introduction

The following equation denotes the derivation of the theoretical values of the aforementioned Summation test (see appendix A).

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

More particularly, the following equation is the distribution of the Summation test evaluated for each digit within the confines of a decade.

2) Distribution = $\frac{N\int_{d10^k}^{(d+1)10^k} \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\,dx}{N\int_{10^k}^{10^{k+1}} \frac{e^{-(\ln(x)-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\,dx}$ d= 1,2,3,4,5,6,7,8,9

The distribution converges to a Benford, and not a Uniform distribution as the standard deviation approaches infinity. The results were computed on an Excel spreadsheet applying numerical integration. This comprises the theoretical values.

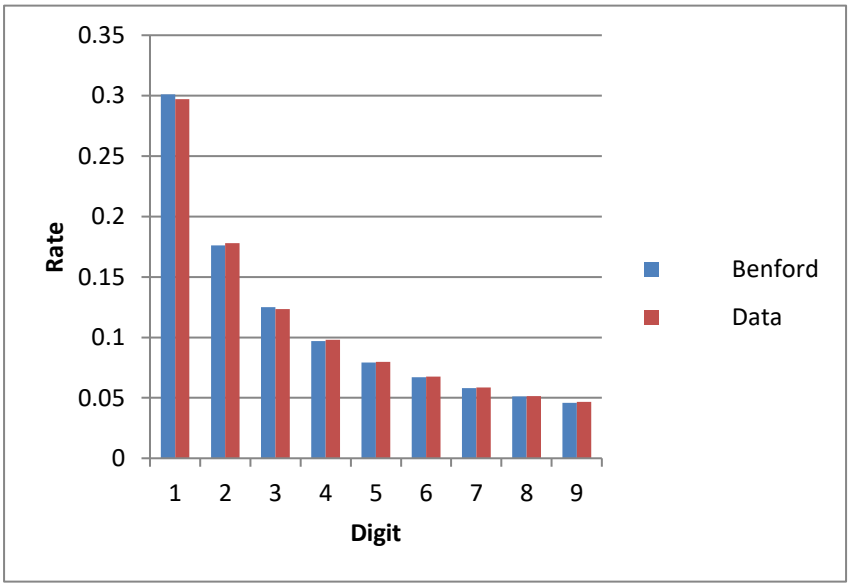
The empirical values were derived from utilizing the Excel Log Normal Inv function which generates random numbers that conform to a Log Normal distribution for a particular mean and standard deviation. I also utilized an Excel spreadsheet with a Visual Basic macro written by myself in order to perform the Summation test as well as other relevant Benford Law tests. I had previously compared the outputs with the corresponding output counterparts obtained from Dr Mark Nigrini's book,"Benford's Law" and achieved identical results.

The following graphs illustrate the Summation test as the standard deviation approaches zero (mean = 2;standard deviation = 1.3,1,0.8,0.6,0.4,0.2) Also, a comparison of the aforementioned theoretical values with the previously defined empirical values.

Fig#1

Summation Test: Log Normal mean = 2; Std Dev = 1.3

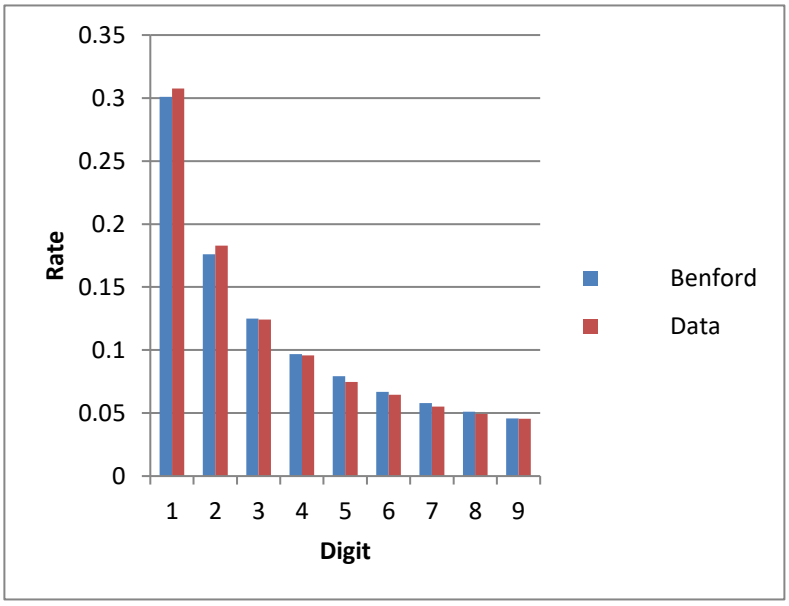
Digit	Benford	Empirical Data
1	0.301029996	0.29703382
2	0.176091259	0.1778865
3	0.124938737	0.12340098
4	0.096910013	0.09790541
5	0.079181246	0.07973427
6	0.06694679	0.06748195
7	0.057991947	0.05855461
8	0.051152522	0.0514611
9	0.045757491	0.0465414



Fig#2

Summation Test: Log Normal mean = 2; Std Dev = 1.0

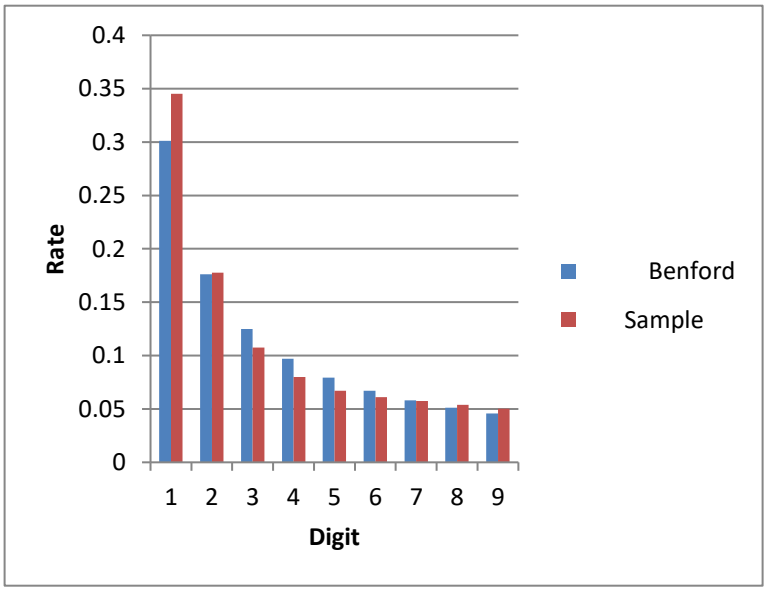
Digit	Benford	Empirical Data
1	0.301029996	0.3075942
2	0.176091259	0.18281784
3	0.124938737	0.12433786
4	0.096910013	0.09580774
5	0.079181246	0.07476766
6	0.06694679	0.06452856
7	0.057991947	0.05525942
8	0.051152522	0.0492929
9	0.045757491	0.0455939



Fig#3

Summation Test: Log-Normal: mean=2; Std Dev=0.8

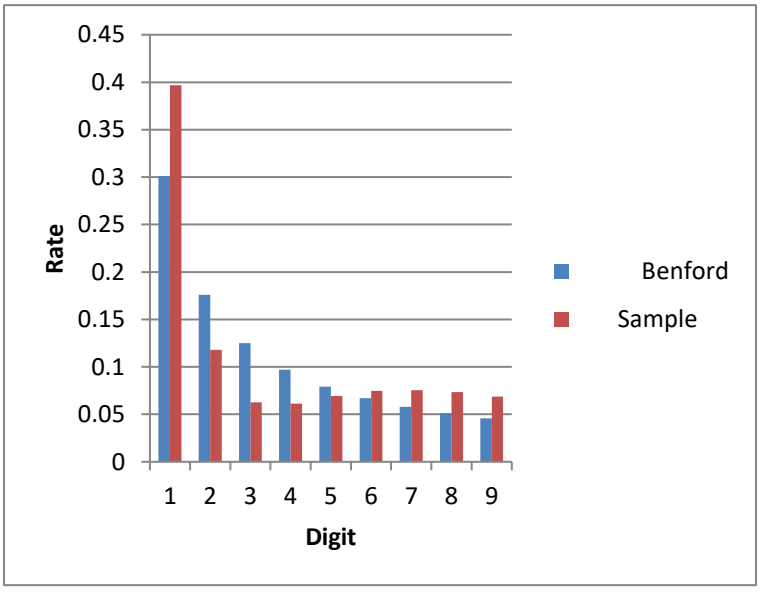
Digit	Benford	Empirical Data
1	0.301029996	0.34521171
2	0.176091259	0.17749406
3	0.124938737	0.10758829
4	0.096910013	0.08005373
5	0.079181246	0.06718732
6	0.06694679	0.06099257
7	0.057991947	0.05733066
8	0.051152522	0.0539293
9	0.045757491	0.0502124



Fig#4

Summation Test: Log Normal: mean=2;Std Dev=0.6

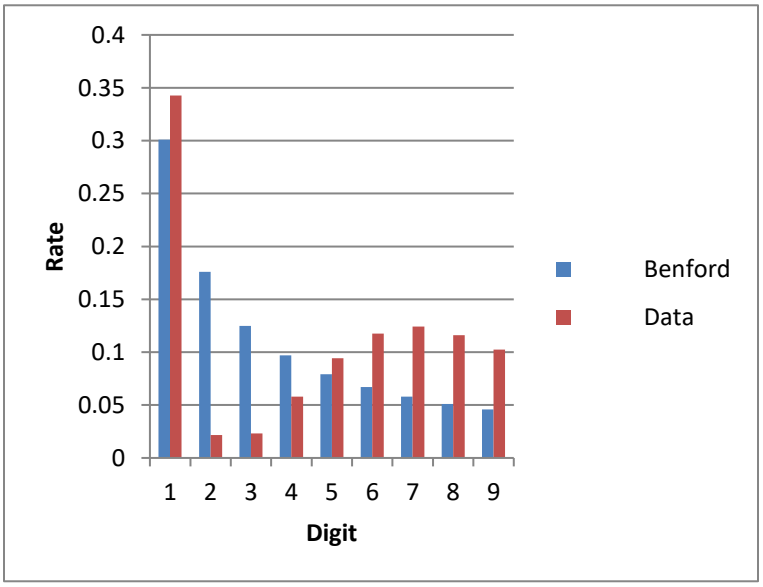
Digit	Benford	Empirical Data
1	0.301029996	0.3967734
2	0.176091259	0.11786601
3	0.124938737	0.06267083
4	0.096910013	0.061262636
5	0.079181246	0.06938147
6	0.06694679	0.07462789
7	0.057991947	0.07556716
8	0.051152522	0.0732963
9	0.045757491	0.0685533



Fig#5

Summation Test: Log Normal mean = 2; Std Dev = 0.4

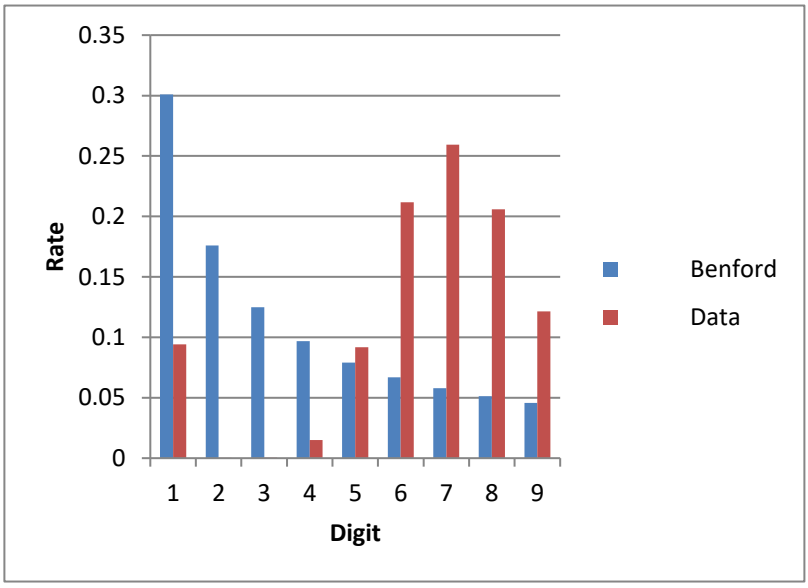
Digit	Benford	Empirical Data
1	0.301029996	0.34277611
2	0.176091259	0.02151943
3	0.124938737	0.023257191
4	0.096910013	0.05797644
5	0.079181246	0.09419991
6	0.06694679	0.11751399
7	0.057991947	0.1242335
8	0.051152522	0.1160908
9	0.045757491	0.1024326



Fig#6

Summation Test: Log Normal mean = 2; Std Dev = 0.2

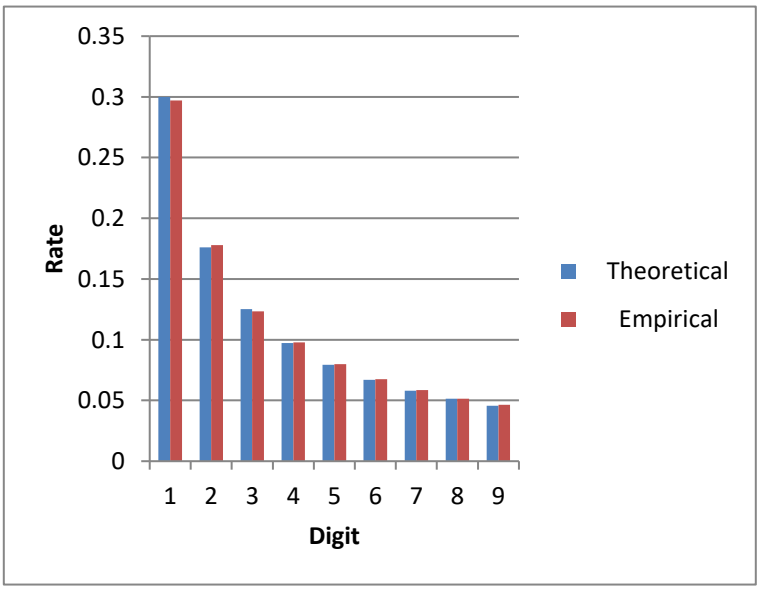
Digit	Benford	Empirical Data
1	0.301029996	0.09408502
2	0.176091259	5.2116E-06
3	0.124938737	0.00053291
4	0.096910013	0.0150587
5	0.079181246	0.09179664
6	0.06694679	0.2117061
7	0.057991947	0.25944798
8	0.051152522	0.2058115
9	0.045757491	0.1215559



Fig#7

Summation Test: mean = 2.0;Std Dev = 1.3

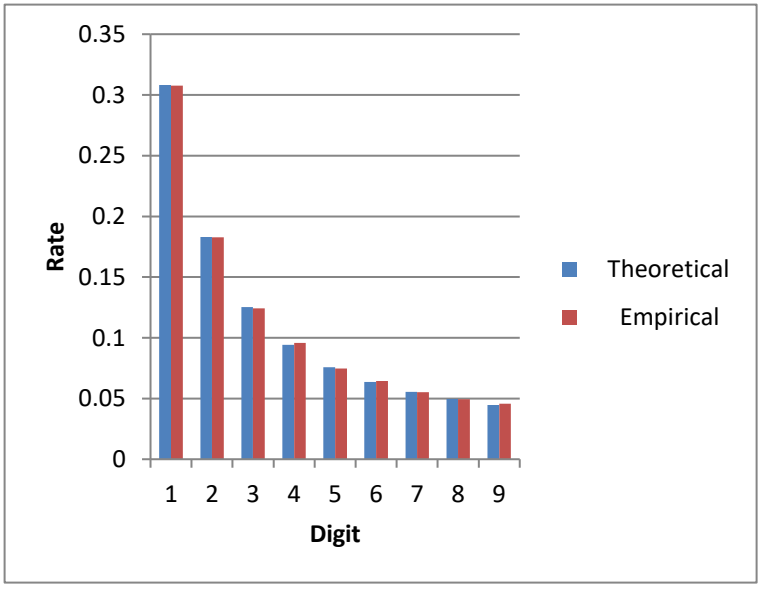
Digit	Theoretical	Empirical
1	0.3000179	0.29703382
2	0.176173859	0.1778865
3	0.125311	0.12340098
4	0.097213251	0.097790541
5	0.0793445	0.079973427
6	0.066987129	0.06748195
7	0.057944561	0.05855461
8	0.051383267	0.0514611
9	0.045624564	0.0465414



Fig#8

Summation Test: mean = 2.0;Std Dev = 1.0

Digit	Theoretical	Empirical
1	0.3082886	0.3075942
2	0.18296869	0.18281784
3	0.1253467	0.12433786
4	0.094186646	0.09580774
5	0.07565123	0.07476766
6	0.06371587	0.06452856
7	0.055470216	0.05525942
8	0.049614203	0.0492929
9	0.044754958	0.0455939

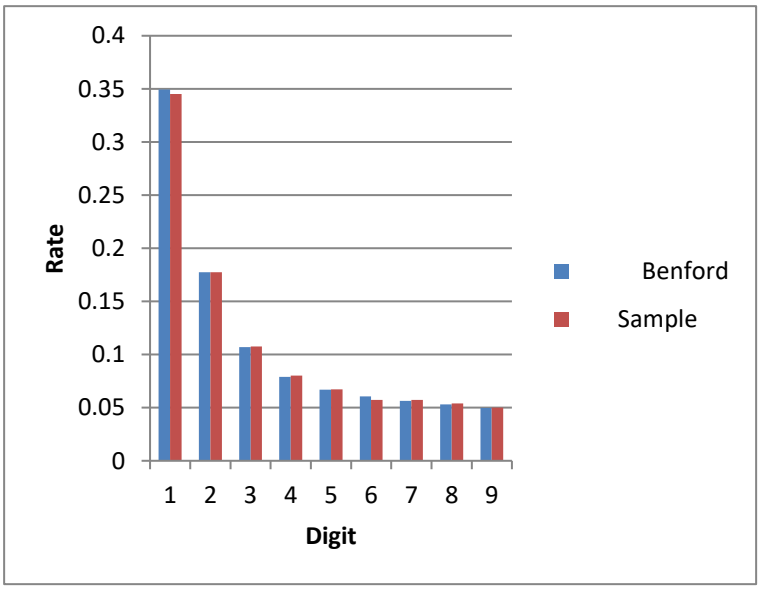


Fig#9

Summation test: mean = 2.0;Std Dev = 0.8

Digit	Theoretical	Empirical
1	0.3492957	0.34521171
2	0.177454394	0.17749406
3	0.1070803	0.10758829
4	0.079076546	0.08005373
5	0.066924042	0.067187321
6	0.060641527	0.05733066
7	0.056489254	0.05733066
8	0.053133117	0.0539293
9	0.049905074	0.0502124

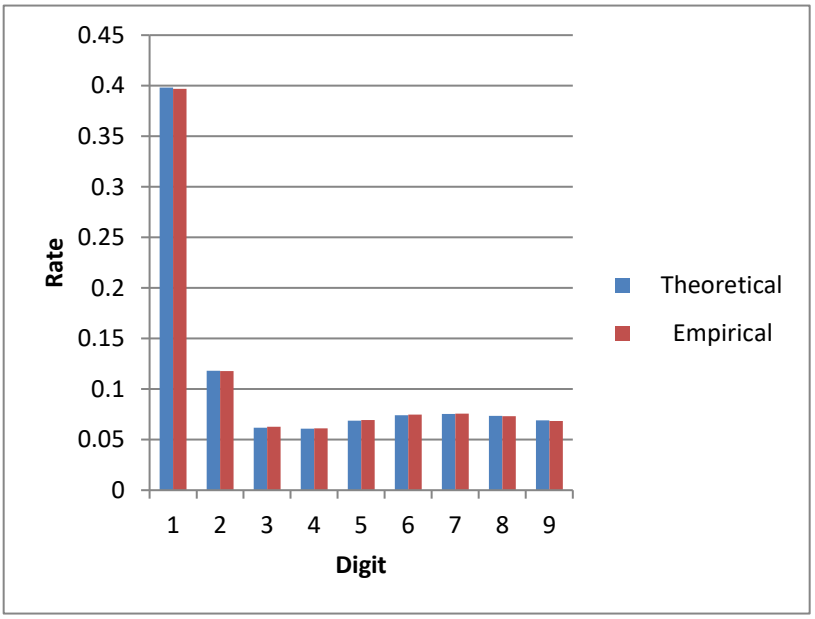
Total



Fig#10

Summation Test: mean = 2.0;Std Dev = 0.6

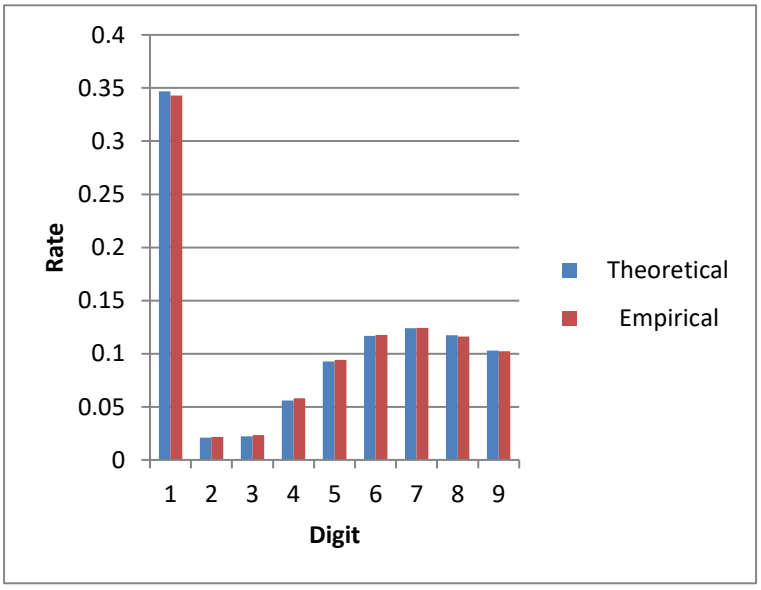
Digit	Theoretical	Empirical
1	0.3982622	0.3967734
2	0.118206718	0.11786601
3	0.0616649	0.06267083
4	0.060939256	0.06126364
5	0.068757581	0.069381471
6	0.074144492	0.07462789
7	0.075453431	0.075567161
8	0.073415853	0.0732963
9	0.069155563	0.0685533



Fig#11

Summation Test: mean = 2.0;Std Dev = 0.4

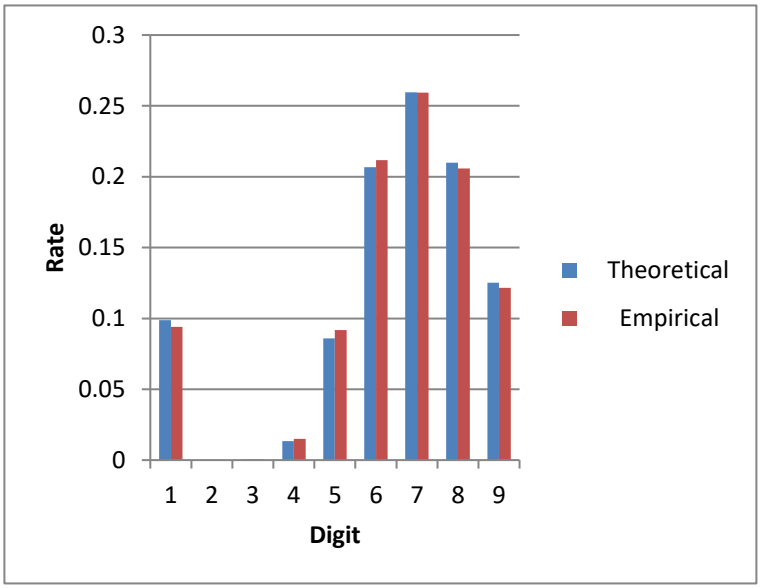
Digit	Theoretical	Empirical
1	0.3469147	0.34277611
2	0.02106246	0.02151943
3	0.0220483	0.023257191
4	0.055921838	0.05797644
5	0.092702094	0.09419991
6	0.116871004	0.11751399
7	0.124000896	0.1242335
8	0.117463357	0.1160908
9	0.103015318	0.1024326



Fig#12

Summation Test: mean = 2.0;Std Dev = 0.2

Digit	Theoretical	Empirical
1	0.0988687	0.09408502
2	0.000017457	5.2116E-06
3	0.000428	0.00053291
4	0.013336401	0.0150587
5	0.085917326	0.09179664
6	0.206620251	0.2117061
7	0.25958277	0.25944798
8	0.210001884	0.2058115
9	0.125238463	0.1215559



Conclusion

It is obvious that the Summation test performed on a data set that conforms to a Log Normal distribution is **not** a Uniform distribution but rather a Benford distribution as the standard deviation approaches infinity. Also, as the standard deviation approaches zero, the Summation test does not result in a Uniform distribution in any way, shape, or form. The Log Normal distribution converges to a Normal distribution with a mean of e^{u} (u equals the mean of the Log Normal distribution and the standard deviation equals the Log Normal standard deviation times the Normal mean – see appendix B).

Case closed

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 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$\int_{\ln(a)-m}^{\ln(b)-m} \frac{e^{-u^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} * e^{u} * e^{m}$ 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 $\sigma \to \infty$ 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. = $\frac{\int_{Ln(d_1)}^{\ln(d_2)} du + \int_{\ln(d_10)}^{\ln(d_20)} du + \int_{\ln(d_100)}^{\ln(d_200)} du + \cdots + \int_{\ln(d_1*10^k)}^{\ln(d_2*10^k)} du}{\int_{Ln(1)}^{\ln(10)} du + \int_{\ln(10)}^{\ln(100)} du + \int_{\ln(100)}^{\ln(1000)} du + \cdots + \int_{\ln(1*10^k)}^{\ln(2*10^{k+1})} du} = \frac{k*\ln(\frac{d_2}{d_1})}{k*\ln(10)} =$ = $LOG_{10}(\frac{d_2}{d_1})$

Appendix B

Proof that as the Standard Deviation of a Log Normal Distribution Approaches Zero 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.

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

Abstract

While it is fairly easy to prove that the Log Normal distribution becomes a Benford distribution as the standard deviation approaches infinity (see appendix A), it is a bit more difficult to prove that as the standard deviation approaches zero that 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.

Proof:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

17) u_x = mean(x); $\boldsymbol{\sigma}_x$ = std_dev(x)

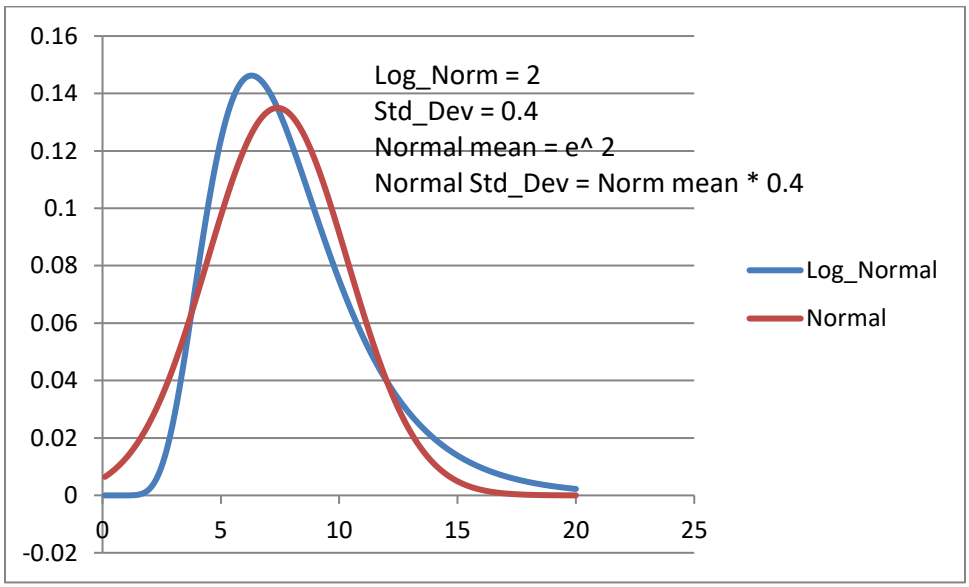
18) $u_x \sim e^u$; $\boldsymbol{\sigma}_x \sim u_x\,\sigma$

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

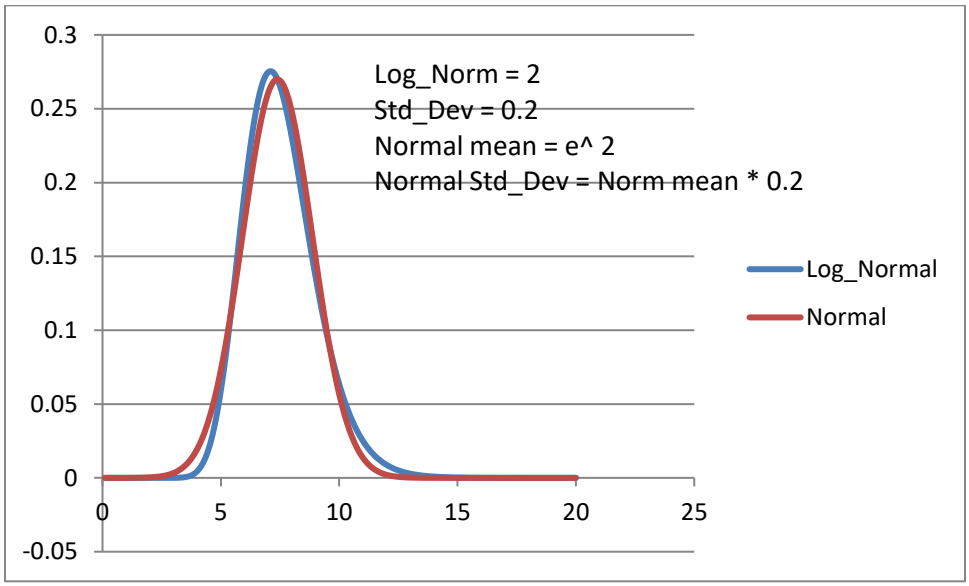
20) Which is a Normal Distribution with a mean of e^u and a standard deviation of σ e^u

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

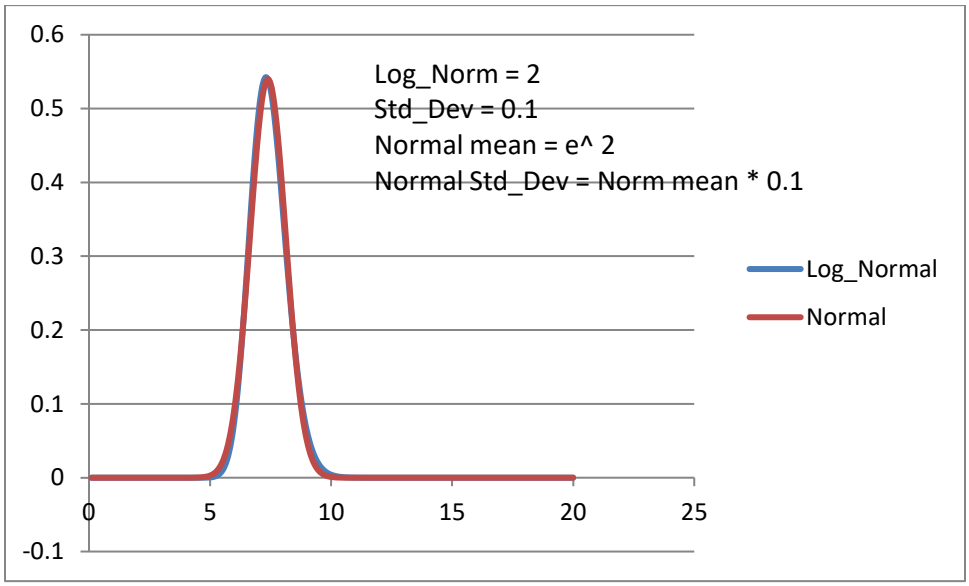
fig#1



fig#2


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#3



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 the mean equal to e raised to Log Normal mean and the standard deviation equal to the Normal times the Log Normal standard deviation.

Appendix C

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 = $\frac{1}{x\sqrt{2\pi\sigma^2}}$ $e^{-(Ln(x)-u)^2/2\sigma^2}$
3) For x=1: 1/xln(10) = 1/ln(10); $\frac{1}{x\sqrt{2\pi\sigma^2}}\, e^{-(Ln(x)-u)^2/2\sigma^2} = \frac{1}{\sqrt{2\pi\sigma^2}}$ $e^{-(u)^2/2\sigma^2}$
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 : $\frac{1}{xln(10)}$ - $\frac{1}{xln(10)}$ ($e^{-(Ln(x)-u)^2/2\sigma^2}$) =

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