

Further Discussion of Benford Coefficients

Lawrence Dworsky

rev. March 10, 2010

Introduction

“Explaining Benford’s Law” is the title of Chapter 34 of the book “The Scientist and Engineer’s Guide to Digital Signal Processing” by Steven W. Smith. This book is available on-line at www.DSPguide.com. Using signal-processing methodology, Dr. Smith clearly and concisely explains the working of Benford’s Law and most importantly explains why distributions which obey Benford’s law show up so often in the world around us.

I highly recommend reading Dr. Smith’s work. Without going through the mathematics, several of the results and observations are

1. The $1/x$ distribution meets all criteria for Benford series numbers, but only when the series is defined over decade (or integer multiple of decade) ranges.
2. The log-normal distribution approximately meets all criteria for Benford series numbers when $\sigma \geq 1$. The larger the standard deviation, the better the approximation. The distribution mean does not affect this situation.
Note that the log-normal distribution, like the normal distribution, does not have a finite upper “edge.” It is, however, always greater than 0.
3. Lists of numbers that occur in “natural” circumstances (e.g. areas of lakes and expense report amounts) often obey Benford statistics because multiplying random distributions leads to the log-normal distribution in the same manner that adding random distributions leads to the normal distribution.

There is more than one way to evaluate how well a set of numbers satisfies Benford’s law. You may either compare the coefficients to the ideal set of Benford coefficients or you may evaluate how constant the coefficients are under arbitrary scaling:

1. Look at the coefficients. There are many ways of calculating a “goodness of fit” for this set of 9 numbers. Without trying to justify this as the best way, I define an RMS error as

$$\sqrt{\sum_{i=1}^9 \left(\frac{b_i}{b_i - B_i} \right)^2} \quad (1)$$

where b_i are the coefficients being considered and B_i are the exact coefficients.

2. Smith defines a function *ost* (“ones scaling test”) by noting the fraction of leading digit 1 in the set, then scaling the set by 1.01 and again noting the fraction of leading digit 1s. This is repeated 696 times, at the end of which the original series has been scaled by a factor of 1,000. There are now 696 numbers which ideally are all equal to $\log 2 = 0.3010$. For a set of numbers which does not obey Benford’s law, these (696) numbers will vary. Since we would expect the same results after scalings of 10, 100 and 1,000, the set of 696 numbers will vary periodically. Looking at the difference between the highest and lowest values in this periodic function gives us a measure of how well the set obeys Benford’s law.

The Log-Normal Distribution

To create some examples, I start with the observation that a set of numbers generated using

$$X = e^{\mu + \sigma N} \tag{2}$$

where N is a set of numbers with a normal distribution, mean = 0 and sigma = 1 will have a log-normal distribution with parameters μ and σ .

Equation 2 may also be written as

$$X = e^\mu e^{\sigma N} \tag{3}$$

The term e^μ may be thought of as a scale factor for the distribution X and as we know, a distribution that does/doesn't obey Benford's law will not lose/acquire the property of obeying Benford's law when multiplied by a scale factor. Therefore, I will set $\mu = 0$ in the examples below.

The histogram for a typical set of 10,000 numbers drawn from a normal distribution with mean 0 and standard deviation is shown in Figure 1.

The leading digit statistics, B_n , for this set of numbers is shown in Table 1.

The rms error = 0.502. The ost function high - low = 0.3885 - 0.2115 = 0.1770. This is clearly a poor fit to the Benford coefficients.

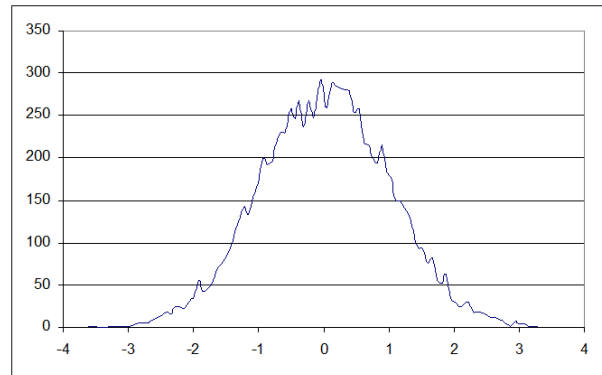


Figure 1: Normal Distribution

Now consider the log-normal distributions generated using the data that generated figure 1 and the formula of equation 2, for $\sigma = 2/3$ and $\sigma = 1$ as shown in figure 2.

For $\sigma = 2/3$, the rms error = 0.649 and the max - min ost values = 0.400 - 0.199 = 0.201; for $\sigma = 1$, the rms error = 0.013 and the max - min ost values = 0.318 - 0.290 = 0.028.

The difference between how two apparently very similar distributions obey Benford's law is striking.

n	B
1	0.36
2	0.13
3	0.09
4	0.08
5	0.08
6	0.07
7	0.07
8	0.07
9	0.06

Table 1: Normal Distribution Leading Digits

Figure 3 shows the dependence of the rms error and the ost values on σ .

Both figures show that, for $\sigma \gtrsim 1$, a data set having a log-normal distribution will obey Benford's law very, very well – albeit never exactly.

Returning to the generation of log-normal distributed data sets by multiplying other distribution data sets together, consider multiplying n 10,000 point data sets which are uniformly distributed between 1 and 100 and then looking at the Benford law conformance.

It doesn't take too many multiplications before the adherence to Benford's law is quite good (Table 2). A caveat here is that there are many parameters involved in this calculation – the lower and upper boundaries of the initial distribution, possibly different initial distributions entirely, and, of course, the number of points in the data set being generated.

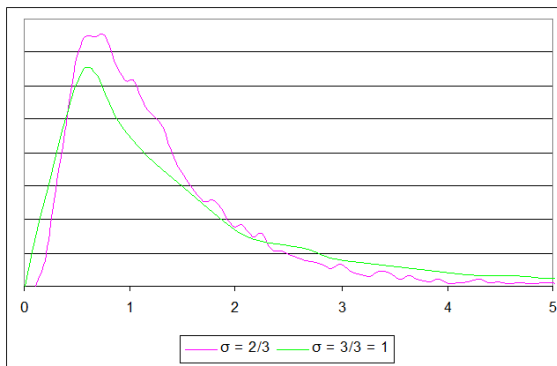


Figure 2: Log-Normal Distributions

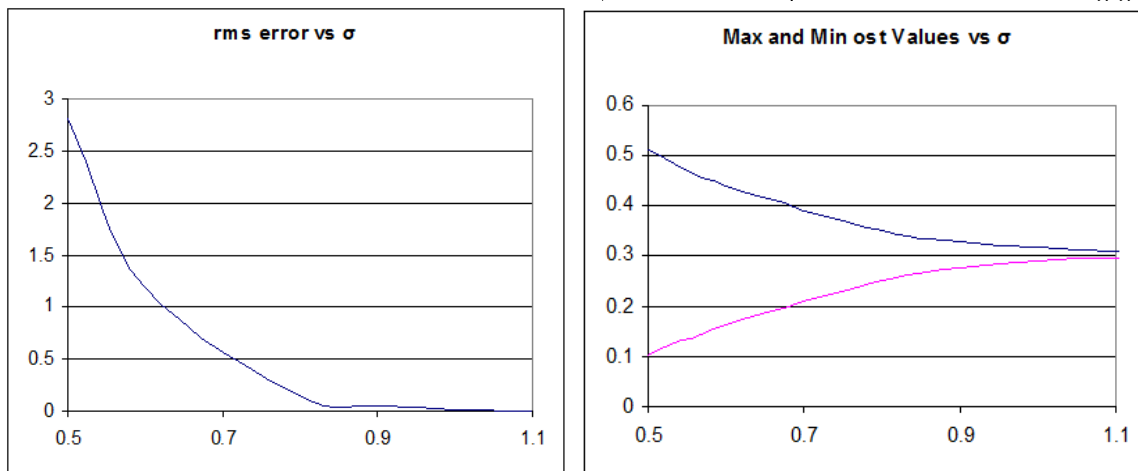


Figure 3: LogNormal Distribution and Benford's Law

The latter factor above affects all examples: Whatever distribution a data set obeys, these are random numbers. There will be 'noise' and the only way to make the data 'look better' is to use larger and larger data sets.

The $1/x$ Distribution

As discussed in chapter 13, a list of random variables obeying the $1/x$ probability distribution function (PDF) will have obey Benford's Law – but only under specific circumstances:

The fraction of a distribution based on a range of random variables $a < x < b$ is found by calculating the number of variables expected from a large set with each leading digit and dividing this number by the total number of variables expected from the same set. For the $1/x$ distribution on $1 < x < 10$ and a leading digit n , the fractional occurrence of the leading digit B_n is found using equation 4.

$$B_n = \frac{\int_n^{n+1} \frac{dx}{x}}{\int_a^b \frac{dx}{x}} \quad (4)$$

n	rms err	ost diff
1	5.56	0.446
2	0.25	0.129
3	0.05	0.049
4	0.02	0.031
5	0.01	0.018

Table 2: LogNormal Distribution by Multiplying

where $a = 1$, $b = 10$ and $1 \leq n \leq 9$.

For example, equation 5 shows the $n = 1$ case

$$B_1 = \frac{\int_1^2 \frac{dx}{x}}{\int_1^{10} \frac{dx}{x}} = \frac{\ln\left(\frac{2}{1}\right)}{\ln\left(\frac{10}{1}\right)} = \frac{\log\left(\frac{2}{1}\right)}{\log\left(\frac{10}{1}\right)} = 0.301 \quad (5)$$

and the entire set is given in table 3.

The basic B_n relationship has some interesting properties. First, as shown in equation 5, the base of the logarithms used doesn't matter.

For the (decade) range $10 < x < 100$, equation 7 shows the same result as equation 5.

$$B_1 = \frac{\int_{10}^{20} \frac{dx}{x}}{\int_{10}^{100} \frac{dx}{x}} = \frac{\log\left(\frac{20}{10}\right)}{\log\left(\frac{100}{10}\right)} = \frac{\log(2)}{\log(10)} = 0.301 \quad (6)$$

n	B
1	0.301
2	0.176
3	0.125
4	0.097
5	0.079
6	0.067
7	0.058
8	0.051
9	0.046

Table 3: Leading Digit Coefficients

Combining both ranges, equation 7 yields

$$B_1 = \frac{\int_1^2 \frac{dx}{x} + \int_{10}^{20} \frac{dx}{x}}{\int_1^{100} \frac{dx}{x}} = \frac{\log\left(\frac{2}{1}\right) + \log\left(\frac{20}{10}\right)}{\log\left(\frac{100}{1}\right)} = \frac{2 \log(2)}{\log(100)} = 0.301 \quad (7)$$

Any number of these decade ranges combined would give the same result. Also, it doesn't matter where the decade range begins. For example, equation 8 calculates finding B_1 in a data set running from 12.5 to 125.

$$B_1 = \frac{\int_{12.5}^{20} \frac{dx}{x} + \int_{100}^{125} \frac{dx}{x}}{\int_{12.5}^{125} \frac{dx}{x}} = \frac{\log 10 + \log 2 - \log 12.5 + \log 10 + \log 12.5 - 2 \log 10}{\log 10 + \log 12.5 - \log 12.5} = 0.301 \quad (8)$$

2nd Digit Distribution

There are Benford statistics for all digits, not just the leading digit. For example, on the range $1 < x < 10$, using the $1/x$ distribution, the population percentage of second digits = 1 is given by equation 9

$$C_1 = \frac{\sum_{k=1}^9 \int_{k+0.1}^{k+0.2} \frac{dx}{x}}{\int_1^{10} \frac{dx}{x}} = 0.114 \quad (9)$$

The full set of second digit coefficients on this same range is shown in table 4.

Comparing tables 3 and 4, you can see that the second digit coefficients are much closer to a constant value (10%) than is the first digit coefficients. Succeeding digits will continue to follow this trend.

n	B
0	0.120
1	0.114
2	0.109
3	0.104
4	0.100
5	0.097
6	0.093
7	0.090
8	0.088
9	0.085

Table 4: 2nd Digit Benford Coefficients