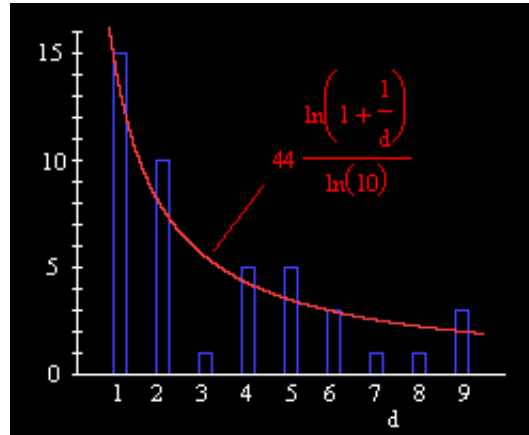## Benford's Law

Benford's Law (which was first mentioned in 1881 by the astronomer Simon Newcomb) states that if we randomly select a number from a table of physical constants or statistical data, the probability that the first digit will be a "1" is about 0.301, rather than 0.1 as we might expect if all digits were equally likely. In general, the "law" says that the probability of the first digit being a "d" is

$$P\{d\} = \frac{\ln\left(1 + \frac{1}{d}\right)}{\ln(10)}$$

This implies that a number in a table of physical constants is more likely to begin with a smaller digit than a larger digit. It was published by Newcomb in a paper entitled "Note on the Frequency of Use of the Different Digits in Natural Numbers", which appeared in The American Journal of Mathematics (1881) 4, 39-40. It was re-discovered by Benford in 1938, and he published an article called "The Law of Anomalous Numbers" in Proc. Amer. Phil. Soc 78, pp 551-72.

To illustrate this interesting fact, try tabulating the first digits of the physical constants listed in Table 2.3 of Abramowitz and Stegun's "Handbook of Mathematical Functions". The result is the bar chart shown below, which gives the distribution of the leading digits of the 44 constants in the table, along with the theoretical expected distribution based on Benford's Law:

Aside from the conspicuous deficiency of 3's, this is a reasonably good match for just 44 data points.

Although there have been many lengthy and erudite "explanations" of Benford's Law, it seems to me it can be explained with a single picture:

1---------------2---------3-------4-----5----6---7--8--9

Clearly the underlying premise of Benford's Law is that the subject population of quantities, expressed in the base 10 and more or less arbitrary units, will be fairly evenly distributed on a logarithmic scale. This is confirmed by the fact that the exponents on these constants are fairly uniformly distributed (at least over several orders of magnitude). As a result, the probability of the leading digit being "d" clearly approaches
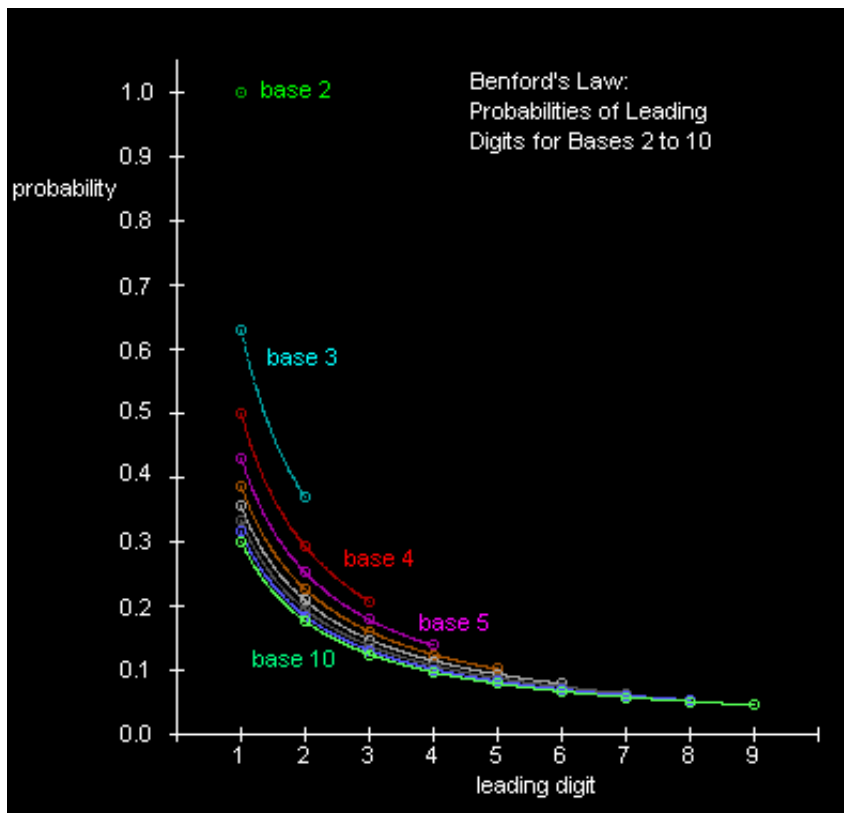
$$\frac{\log_{10}(d+1) - \log_{10}(d)}{\log_{10}(10) - \log_{10}(1)} \quad = \quad \log_{10}\left(1 + \frac{1}{d}\right) \quad = \quad \frac{\ln\left(1 + \frac{1}{d}\right)}{\ln(10)}$$

Of course, we *could* have chosen units for our physical constants such that the leading digits were all 9's (for example), but evidently we have a natural tendency to choose units so that our numbers are evenly distributed by order of magnitude, rather than absolute value. This may be related to our basic impressions of hearing and sight (and earthquakes), since our sense impressions of loudness and brightness are logarithmic.

Naturally we can apply Benford's Law to numbers expressed in any base, not just the base 10. In general the probability of the leading digit d (in the range 1 to B-1) for the base B is

$$P\{d\} \quad = \quad \frac{\ln\left(1 + \frac{1}{d}\right)}{\ln(B)}$$

Notice that for binary numbers, i.e., numbers expressed in the base 2, the probability of the leading digit being 1 is 1.000, as it must be, since the leading non-zero digit of a binary number is necessarily 1. The distributions of probabilities of the digits 1 to B-1 for each base B from 2 to 10 are shown below.

We can also easily verify that the sum of all the probabilities for digits 1 through b-1 equals 1.0000, as it must, since the leading digit must be one of these. This implies
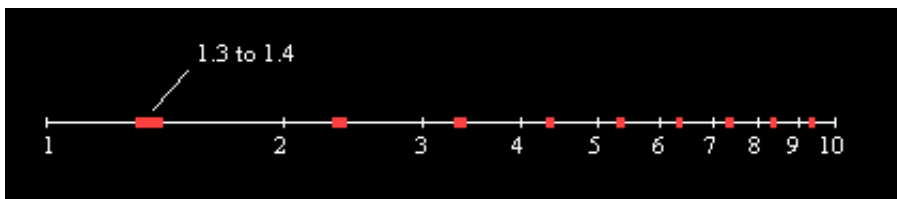
$$\ln\left(1 + \frac{1}{1}\right) \quad + \quad \ln\left(1 + \frac{1}{2}\right) \quad + \quad \ldots \quad + \quad \ln\left(1 + \frac{1}{B-1}\right) \quad = \quad \ln(B)$$

To verify this, recall the fundamental law of logarithms, $\ln(ab) = \ln(a) + \ln(b)$. With this we can re-write this sum of logarithms as the logarithm of a product:

$$\ln\left(\frac{2}{1}\cdot\frac{3}{2}\cdot\frac{4}{3}\cdot\ldots\cdot\frac{B}{B-1}\right) \;=\; \ln(B)$$

which confirms the result.

By the same kind of analysis we can determine the probability that the *second* digit will have a certain value. It's only necessary to consider a single order of magnitude, since the pattern is repeated on each order. For example, in the base 10, the probability of the second digit being "3" is equal to the sum of the probabilities of the first two digits being "1.3", "2.3", "3.3", ... or "9.3" for numbers in the range from 1 to 10. This is indicated by the shaded regions in the logarithmic scale shown below.

The fraction of this region covered by the range from 1.3 to 1.4 is

$$\frac{\ln(1.4) - \ln(1.3)}{\ln(10) - \ln(1)} \;=\; \frac{\ln\left(\frac{14}{13}\right)}{\ln(10)}$$

The fraction covered by the other regions (such as 2.3 to 3.3, and so on) can be found similarly, and we can add them together to give the total probability that the first digit *following* the first non-zero digit will be a 3:

$$P_1\{3\} \;=\; \frac{\ln\left(\frac{14}{13}\right) + \ln\left(\frac{24}{23}\right) + \ldots + \ln\left(\frac{94}{93}\right)}{\ln(10)}$$

In general, the probability of the 2nd digit of d in a base-B number (taken from a logarithmic population) is

$$P_1\{d\} \;=\; \frac{1}{\ln(B)}\sum_{k=1}^{B-1}\ln\left(1+\frac{1}{kB+d}\right) \;=\; \frac{1}{\ln(B)}\ln\left[\prod_{k=1}^{B-1}\left(1+\frac{1}{kB+d}\right)\right]$$

Extending this analysis to the case of the nth digit following the first non-zero digit, we arrive at the general formula

$$P_n\{d\} \;=\; \frac{1}{\ln(B)}\sum_{k=B^{n-1}}^{B^n-1}\ln\left(1+\frac{1}{kB+d}\right)$$

This applies to the case of the leading non-zero digit, with the understanding that with n = 0 the summation reduces to just the single term k = 0. This formula shows that the non-uniformity in the distribution of digits becomes much less as we consider less significant

digits. For example, we have $P_0\{1\} = 0.301019995...$, $P_1\{1\} = 0.113890103...$, and $P_2\{1\} = 0.101375977...$ Thus the probability of a "1" quickly approaches 1/10 as we proceed to less significant digits.