BENFORD'S LAW

J. W. SANDERS

Pick a book containing lots of four digit numbers (for example a telephone directory or book of maths tables), and choose a number, let's call it N, from the book, at random. (Any method you like can be used to choose the number - for instance in the well known cockroach method, you let a cockroach wander onto a page of the book, then close it suddenly; the only problem with this method is that sometimes the chosen number is hard to read ...) Let us call the first non-zero digit in N, counting from the left, the *first significant digit of* N. I bet you that the first significant digit of N is less than or equal to 4. Whoever loses the bet pays the winner \$1; do you take the bet?

Unless you object on moral grounds, you might argue like this: the first significant digit can be anything from 1 to 9, and since these appear equally likely, the probability of getting less than or equal to 4 is exactly 4/9. Since $4/9 < \frac{1}{2}$, you have a better than even chance of winning, so you presumably take the bet. (Warning: herein lies the path to ruin!)

Folklore has it that one applied mathematician coasted to fame and fortune on this bet. In fact experience shows that the probability of the first significant digit of our random number N being ≤ 4 is about 7/10, so you have only probability 3/10 of winning - considerably less than an even chance. Why is this so?

First let's look at the evidence. F. Benford used a variety of collected numbers (data) obtained in some sort of investigation and has compiled the frequency with which a randomly chosen number N has first significant digit equal to each of the digits from 1 to 9. His tabulated frequencies are reproduced entirely (see Table 1). In each row the frequencies appear from different sources (Benford's paper gives no further detail about the nature of the data) and averages of these appear in the bottom row. Observe that if each digit were equally likely, every entry would be close to $\frac{100}{9} = 12.4$.

From the bottom row of table 1 we can read off the experimentally obtained probabilities with which N can be expected to have its different first significant digits - these appear in row (*) of table 2.

Source of numbers	First significant digit								
	1	2	3	4	5	6	7	8	9
Geographical data	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1
Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2
Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6
Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0
Specific heats	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1
Pressures	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7
H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6
Molecular wt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2
Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9
Atomic wt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5
$n^{-1}, \sqrt{n}, \ldots$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9
Design	26.8	14.8	14.3	7.3	8.3	8.4	7.0	7.3	5.6
Digest	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2
Cost data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1
X-ray volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8
U.S. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0
Black body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4
Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0
$n, n^2, \ldots, n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5
Death rates	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1
AVERAGE	30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7

Percentage of times the digits from 1 to 9 appear as the first significant digit of a randomly chosen number N, as given by 20,229 observations.

TABLE 1

	First significant digit									
	1	2	3	4	5	6	7	8	9	
Experimental probabilities	.306	.185	.124	.094	.080	.064	.051	.049	.047	(*)
log(n+1) - log(n)	.301	.176	.125	.097	.079	.067	.058	.051	.046	(**)

Experimental probabilities compared with those calculated using the formula log(n+1) - log(n).

TABLE 2

We can now see why 0.7 equals the probability that the first significant digit of N is ≤ 4. Indeed the probability of getting ≤ 4 is the probability of getting 1, plus the probability of getting 2, plus the probability of getting 3, plus the probability of getting 4. That is,

$$.306 + .185 + .124 + .094 = .709.$$

So our task is to explain the frequencies occurring in row (*) of table 2. We start by observing Benford's Law: if n is the first significant digit of N then the experimentally obtained probability of n occurring (row (*) of table 2) is closely matched by $\log(n+1) - \log(n)$ — these are recorded in row (**) of table 2.

We now have to explain Benford's Law, and argue as follows. The first significant digit of N being k means that, for some power m,

$$k.10^m \leq N < (k+1)10^m \qquad \ldots (1)$$

(e.g. 0534 has first significant digit 5 since $534 = 5.34 \times 10^m$ for $m = 2$). Taking logs, (1) holds if and only if

$$\log(k.10^m) \leq \log N < \log\left((k+1)10^m\right), \qquad \ldots (2)$$

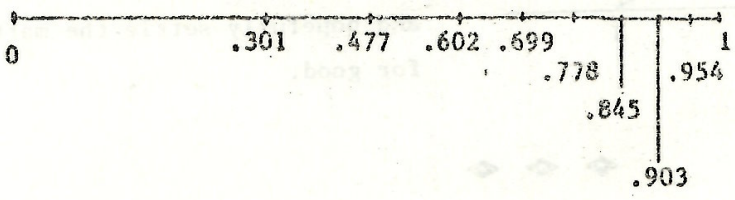
because $x < y$ if and only if $\log x < \log y$. Next by properties of logs, (2) holds if and only if

$$\log(k) + m \leq \log N < \log(k+1) + m. \qquad \ldots (3)$$

So the probability of (1) occurring equals the probability that (3) occurs - i.e. the probability that $\log N$ lies in the interval $[\log(k) + m, \log(k+1) + m)$. We might expect this probability to be proportional to the length of the interval, namely

$$\log(k+1) - \log(k). \qquad \ldots (4)$$

For $k = 1,2, \ldots,9$ these lengths are

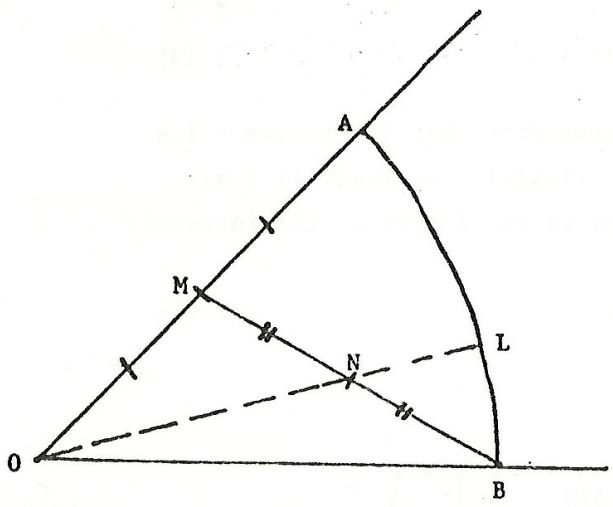


and since they fill out the interval [0,1], the constant of proportionality must be 1. In other words (4) is the probability with which N has first significant digit k.

How convincing is this justification of Benford's Law? Why should the probability of (3) be proportional to the length of the interval; i.e. why should log N be evenly distributed? After all, we are trying to explain why N is *not* evenly distributed. This is really an *assumption* about N, whose only advantage is that it leads to the correct answer.

There is a justification of Benford's Law which surmounts this difficulty, and can be summarised: suppose that, whatever the probability with which N takes on first significant digit k is, it is unchanged when all the numbers in the book undergo a scale change. Then the probability of getting k *must be* log(k + 1) - log(k). The proof, due to Pinkham, is far harder than the one above, and will be sketched in a later article.

TRISECTION OF THE ANGLE?



One of our readers, James Thorpe (Year 10) sent us this construction. ∠AOB is to be trisected. Construct: arc AB of any radius, bisect OA at M, then MB at N. The ∠LOB is the third (?) of ∠AOB.
Can you prove it or can you find the fallacy?
In our next issue we will dig deeper into this ancient problem and hopefully settle the matter for good.