THE LAW OF ANOMALOUS NUMBERS

FRANK BENFORD

Physicist, Research Laboratory, General Electric Company,
Schenectady, New York

PROCEEDINGS OF THE AMERICAN PHILOSOPHICAL SOCIETY,
VOL. 78, NO. 4, MARCH, 1938

Frank Benford introduced what is now called “Benford’s Law” in a paper in 1938 in which he collected a large amount of heterogeneous data and tracked how often each possible first digit (1,2,3,..,9) appeared among the numbers. The data included 1458 randomly selected baseball figures, the areas of 335 rivers, 342 street addresses, 308 numbers pulled from Reader’s Digest, etc. etc.

TABLE I

PERCENTAGE OF TIMES THE NATURAL NUMBERS 1 TO 9 ARE USED AS FIRST DIGITS IN NUMBERS, AS DETERMINED BY 20,229 OBSERVATIONS

Group	Title	First Digit									Count
		1	2	3	4	5	6	7	8	9	
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	$n^{-1}, \sqrt{n}, \cdots$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	*Digest*	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	$n^1, n^2 \cdots n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average.....		30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Probable Error		±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.3	—

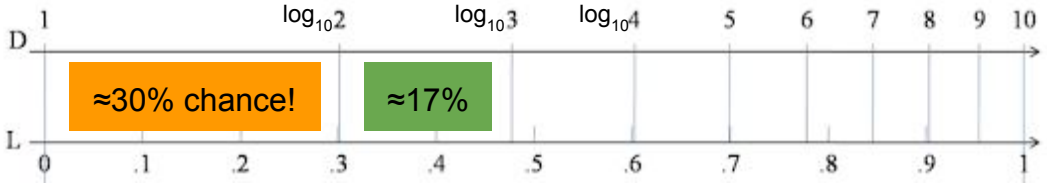
He found that rather than all first digits being equally likely (which would translate to 11.111…% frequency), there was a pronounced bias toward lower numbers. In particular, over 30% of the numbers in his sample begin with a 1, while under 5% begin with a 9.

What's going on here? The idea is that data in the world are collected on many *scales*—measurements are made in different units, numbers like street addresses range over different scales (like if one street is numbered from 1-500 and another street is numbered 1-250).

So we shouldn't expect the data to be uniform; rather, we should expect the histogram to have roughly the same shape if we rescale the data, like by the function f(x)=2x. (This corresponds to a change of units or a change of scale.)

This suggests that logarithms would be a useful way to study the data, because log(2x)=log 2 + log x, so if the original data is *scale-invariant*, we'd expect the log data to be *translation-invariant*. (Let's take all our logs to be base 10 so they are compatible with place value notation.)

If we write a number N in scientific notation $N=a\cdot 10^m$, then the first digit of N equals 1 if and only if $1 \le a < 2$, which happens if $0 \le \log a < \log 2$. But log 2=.301… Thus if the (log a) values are uniformly distributed over the possibilities from 0 to 1, we should get first digit 1 about 30.1% of the time!



Benford's Law

Leading Digit	Occurring Frequency
1	30.10%
2	17.60%
3	12.50%
4	9.70%
5	7.90%
6	6.70%
7	5.80%
8	5.10%
9	4.60%