The first-digit law

Antonella Perucca

Nowadays, in times of information explosion, we are surrounded by huge amount of data that we have to handle. The data seems chaotic, but often it obeys some simple statistical and probability rules.

Benford's law, also called *first-digit law*, states that the frequencies of the first digits in many real-life sets of numerical data are quite regular, being roughly the following:

First digit	1	2	3	4	5	6	7	8	9
Frequency (percentage)	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6

The above percentages are roundings, the exact probability for a digit d between 1 and 9 is:

$$P(d) = \log_{10}(d+1) - \log_{10}(d) = \log\left(1 + \frac{1}{d}\right).$$

Notice that by 'first digit' we mean the first non-zero digit, so for example the first digit of the number $0,0123$ is 1.

Benford's law is surprising: why should the digit 1 be much more frequent than the other digits? One partial explanation is that the numbers with first digit 1 come by counting before numbers with other digits and therefore appear more often in practice.

It is also amazing that if a set of data satisfies Benford's law while expressed in some unit, then after a rescaling (i.e. after multiplication with a constant) the data still satisfies the law. This is obvious if we multiply the data by a power of 10, but the property holds for any constant. Let us exemplify this in an easy case [1], supposing that we halve the unit, so that we have to multiply the data by 2. Then we have:

First digit of n	1	2	3	4	5 or 6 or 7 or 8 or 9
First digit of $2n$	2 or 3	4 or 5	6 or 7	8 or 9	1

Since

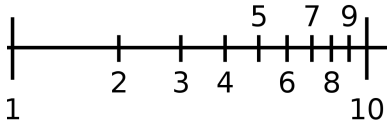
$$P(5) + P(6) + P(7) + P(8) + P(9) = P(1)$$

we deduce that the frequency for the digit 1 of the number $2n$ is, as it should be, equal to $P(1)$.

Benford's law as a mathematical statement (which has a proof) is the following [1]:

> If you take random samples from randomly chosen data sets, then the more set and samples you select, the closer and closer the probability of having first digit d will be to $\log\left(1+\frac{1}{d}\right)$.

In a stronger version, Benford's law states that *the fractional part of the decimal logarithm* of the values are uniformly distributed. Notice that a number starts with a digit d between 1 and 9 if and only if the fractional part of its decimal logarithm is between $\log_{10}(d+1)$ and $\log_{10}(d)$ (to see this, rescale the values by a power of 10, so that they are between 1 and 10). Because of the uniform distribution, the probability for a value to have first digit d is proportional to $P(d) = \log_{10}(d+1) - \log_{10}(d)$, which is the length between $d+1$ and d represented on the logarithmic scale:



Since the sum of the $P(d)$'s is $\log(10) - \log(1) = 1$, the probability is in fact equal to $P(d)$.

Benford's law applies to mathematical objects, for example if we look at more and more terms of some sequences, then the first digits are distributed according to this law. Among such sequences there is the sequence of the factorials (namely, the sequence $n!$), of the powers of 2 (namely, the sequence 2^n), and of the Fibonacci numbers (namely, the sequence starting with 0 and 1 and such that the further terms equal the sum of their two previous terms).

Benford's law applies best in practice to set of datas which look like random and which stretch over several order of magnitudes. It does not apply to all sets of data. To name a simple example, if we measure in degree Celsius the human temperature, then the first digit of the values (which are around $40°$) will be 3 or 4. Nevertheless, Benford's law applies quite well to various kinds of financial data: it has even been possible to discover some financial frauds because suspicion arose from the fact that the falsified data did not comply with Benford's law.

References

[1] Alex Bellos, *Alex through the looking glass*, Bloomsbury Publishing, 2014.

[2] Wikipedia contributors. *Benford's law.* Wikipedia, The Free Encyclopedia, `https://en.wikipedia.org/wiki/Benford%27s_law`, retrieved December 20, 2020.

Exercises for the reader

1. Check that $P(5) + P(6) + P(7) + P(8) + P(9) = P(1)$.

2. Can you state Benford's law in a basis b different from 10?

3. Can you state Benford's law for a digit other than the first one?
   Hint: Consider a natural number n and its string of digits, and write down a very simple formula, much similar to that of Benford's law, expressing the probability that a number starts with the digits of n (possibly after initial digits all equal to 0).

Solutions to the exercises for the reader

1. By applying the definition of $P(d)$ we get

$$P(5) + P(6) + P(7) + P(8) + P(9) = \log_{10}(10) - \log_{10}(5)$$

$$= \log_{10}(2) = \log_{10}(2) - \log_{10}(1) = P(1)\,.$$

2. The probability for the first digit in base b to be equal to d, for d varying from 1 to $b-1$, is

$$P(d) = \log_b(d+1) - \log_b(d)\,.$$

3. The probability for the first digits to be those of a natural number n is

$$P(n) = \log_{10}(n+1) - \log_{10}(n) = \log_{10}\left(1 + \frac{1}{n}\right).$$

Thus the probability that a fixed digit d from 0 to 9 is encountered as the m-th digit for some fixed $m \geq 2$ is, with the summation symbol,

$$\sum_{k=10^{m-2}}^{10^{m-1}-1} \log_{10}\left(1 + \frac{1}{10k+d}\right).$$

This can be seen by considering all natural numbers of the form $n = 10k + d$ with k varying from 10^{m-2} to $10^{m-1} - 1$, which give all numbers with m digits having the last digit equal to d.