Derivation of Benford's Law—C.E. Mungan, Fall 2021

*reference: AJP **89**, 851 (2021)*

Suppose a list is made of the lengths L_n (where n = 1, 2, 3, …) of the rivers in the continental United States. One river might have a length of 381 miles, another a length of 1047 miles, and so on. The first digits of these two rivers are 3 and 1, respectively. There are 9 different possibilities for what the first digit could be (namely 1 to 9). What is the probability that a randomly chosen first digit in this list of lengths has a particular value? Surprisingly, the answer is *not* 1 out of 9! It is most likely the first digit will be a 1, markedly less likely it will be a 2, and the probability monotonically decreases thereafter such that the least likely value for the first digit is a 9. Of course, there is nothing special about river lengths that causes this phenomenon. It is a quite general result for any list of measured values and is known as Benford's law.

To prove it, start by writing the lengths in scientific notation as $L_n = x_n \times 10^{k_n}$. For example, 1047 mi = 1.047×10^3 mi so that $x_n = 1.047$ mi and $k_n = 3$. We can assume that $1 \le x_n < 10$. Let $P dx$ be the probability (with probability density P which in the current example of river lengths is in units of mi^{-1}) that x_n lies between the values x and $x + dx$. The surprise is that P is not a constant but is instead a function of x. We must normalize the probability, such that

$$\int_1^{10} P\,dx = 1. \tag{1}$$

The key insight that explains Benford's law is that we require P to be *invariant under a change of scale*. In other words, if we change the units of the lengths from miles to kilometers say, we don't want the first-digit probabilities to change. But a bit of thought will lead one to the realization that that implies a first digit of 1 must be much more likely than a first digit of 9. For example, if the change in scale leads to a doubling of the lengths in the new system of units, then any lengths that previously were in the range from 500 to 999 (and thus had first digits between 5 and 9) will now have lengths in the range from 1000 to 1998 (which *all* start with a 1).

In general, suppose the scale factor is λ. For instance, to convert x_n in miles to λx_n in kilometers, we have

$$\lambda = 5280 \tfrac{\text{ft}}{\text{mi}} \times 12 \tfrac{\text{in}}{\text{ft}} \times 2.54 \tfrac{\text{cm}}{\text{in}} \times 0.01 \tfrac{\text{m}}{\text{cm}} \times 0.001 \tfrac{\text{km}}{\text{m}} \approx 1.609 \tfrac{\text{km}}{\text{mi}}. \tag{2}$$

The scale invariance mathematically requires that the probability $P(\lambda x)d(\lambda x)$ that λx_n lies between λx and $\lambda x + d(\lambda x)$ in units of kilometers is related to the probability $P(x)dx$ that x_n lies between x and $x + dx$ in units of miles according to

$$P(\lambda x) = \lambda^{-1} P(x) \tag{3}$$

because that is the only way to ensure it is properly normalized such that

$$\int_1^{10} P(\lambda x) d(\lambda x) = \int_1^{10} \lambda^{-1} P(x) \lambda\, dx = 1 \tag{4}$$

using Eq. (1) in the last step.

Now differentiate both sides of Eq. (3) with respect to λ to get

$$\frac{d}{d\lambda}P(\lambda x) = -\lambda^{-2}P(x). \tag{5}$$

However, the left-hand side of this result can be rewritten as

$$x\frac{d}{d(\lambda x)}P(\lambda x) \equiv xP'(\lambda x) \tag{6}$$

so that Eq. (5) becomes

$$xP'(\lambda x) = -\lambda^{-2}P(x). \tag{7}$$

Now substitute $\lambda = 1$ into this equation to get

$$xP'(x) = -P(x) \tag{8}$$

which is a differential equation whose solution by inspection is

$$P(x) = \frac{1}{x\ln 10} \tag{9}$$

where the factor of ln10 comes from the normalization condition in Eq. (1).

Finally, the probability that the first digit is d (where d is an integer between 1 and 9) is

$$p_d \equiv \int_d^{d+1} P\,dx = \frac{1}{\ln 10}\ln\left(1+\frac{1}{d}\right) = \boxed{\log\left(1+\frac{1}{d}\right)} \tag{10}$$

which is Benford's law. The last step comes from changing the base of the logarithm from e to 10, as follows from

$$10^{p_d} = 10^{\frac{1}{\ln 10}\ln\left(1+\frac{1}{d}\right)} = \left(e^{\ln 10}\right)^{\frac{1}{\ln 10}\ln\left(1+\frac{1}{d}\right)} = e^{\ln\left(1+\frac{1}{d}\right)} = 1+\frac{1}{d}. \tag{11}$$

The following is a table of the numerical values for the probabilities for the first digits.

d	p_d
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%

The sum of the nine values of p_d is 100% as expected. Thus it is more than six times as likely that a river length will start with a 1 than that it will start with a 9.