Physics Recreations: Benford's Law

Dr. D.G. Simpson
Department of Physical Sciences and Engineering
Prince George's Community College

October 21, 2009

Occasionally we run into problems that are baffling because they don't seem to provide enough information for their solution. For example, look at the data in Table 1, which shows the per capita annual energy use for 32 different countries. Two columns of data are shown in the table: one of the columns is real data, and the other is fake data. The question is: which column shows the real data, A or B?

At first glance, it appears that there is no possible way to tell which column contains the real data—we simply aren't given enough information. But it turns out that there is a way to distinguish the real data from the fake data, without knowing anything about the countries listed.

In the 1880s, astronomer Simon Newcomb noticed that when he went to the library to use their table of logarithms, the pages near the front of the book were much more worn than those toward the end. After doing some investigating, he discovered that in many cases experimental data has the property that the leading digit is 1 most often, the leading digit is 2 somewhat less often, and so on—with the leading digit being 9 least often. This principle was re-discovered by physicist Frank Benford in the 1930s, and is known as *Benford's Law*. It states that in many situations, real-world data has the leading digit d with probability

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

Computing this for each of the digits $1 \ldots 9$, we find

d	$P(d)$
1	0.3010
2	0.1761
3	0.1249
4	0.0969
5	0.0792
6	0.0669
7	0.0580
8	0.0512
9	0.0458

so that the leading digit should be 1 about 30% of the time, but it should be 9 less than 5% of the time.

Table 1. Per capita energy usage by country (in GJ/year).

Country	Column A	Column B
Albania	28.29	67.02
Belgium	239.54	199.75
Brazil	44.84	86.27
Bulgaria	105.34	61.62
Canada	348.63	342.93
China	47.81	64.78
Columbia	26.75	56.43
Egypt	31.97	47.61
France	189.77	255.96
Ghana	16.81	30.54
Greece	113.34	151.54
Guatemala	25.53	36.29
Iceland	492.16	697.20
India	21.52	47.37
Italy	131.34	85.57
Jamaica	64.89	99.71
Japan	169.70	76.41
Kenya	20.21	55.64
Luxembourg	395.17	509.80
Nepal	14.11	47.56
Norway	249.21	369.28
Portugal	104.24	74.87
Qatar	898.62	754.22
Russia	185.77	84.29
South Africa	109.07	80.44
Tajikistan	21.05	50.74
Togo	18.70	36.16
Turkey	46.44	47.39
United Kingdom	164.56	75.92
United States	327.38	530.31
Yemen	12.38	37.21
Zambia	25.23	10.11

Now let's look back at the data in Table 1. If we count up the number of times the leading digit in column A is $1, 2, \ldots, 9$, then do the same for column B, we find:

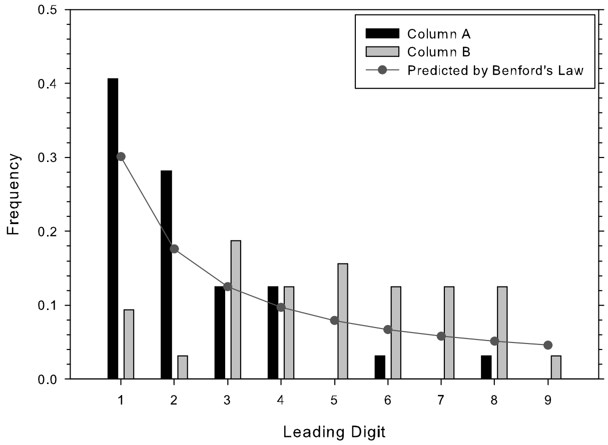
d	Column A	Column B
1	13	3
2	9	1
3	4	6
4	4	4
5	0	5
6	1	4
7	0	4
8	1	4
9	0	1

Dividing each of these by the number of countries (32) gives the frequency distributions of the leading digits:

d	Column A	Column B
1	0.4063	0.0938
2	0.2813	0.0313
3	0.1250	0.1875
4	0.1250	0.1250
5	0.0000	0.1563
6	0.0313	0.1250
7	0.0000	0.1250
8	0.0313	0.1250
9	0.0000	0.0313

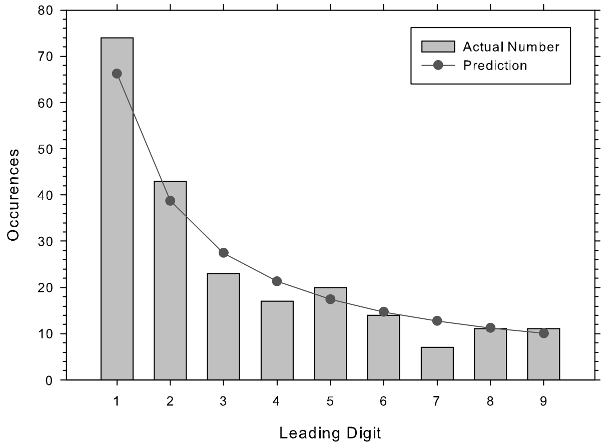
These frequencies are plotted in Figure 1, along with the frequencies predicted by Benford's Law. Clearly column A is the real data, since it follows Benford's Law more closely; the fake data is column B.

Fig. 1 Leading Digit
Frequencies for Data in Table 1



The same situations applies in many other cases. For example, you can check the leading digits in the tables of physical constants you received at the beginning of the semester. The results are shown in Figure 2. (There is better agreement with Benford's Law here than in the previous example because there are more numbers, and therefore better statistics.)

Fig. 2 Benford's Law Applied to the Fundamental Physical Constants (CODATA 2006)



Benford's Law is quite counter-intuitive, since you would probably expect the leading digit of numbers in a data set to be each of the digits 1 through 9 with equal probability. The reason most numbers begin with 1 has to do with the way the data is distributed: if the data is distributed by an *exponential distribution*, then the leading digits will be equally distributed *logarithmically*, so that the leading digit 1 occurs most frequently. Not all data is distributed this way, though. You could not, for example, apply Benford's Law to data on the heights of human adults. If the heights are measured in feet, there will be a lot of numbers whose leading digit is 5, a smaller number beginning with 4 or 6, and few or no numbers whose leading digit is 1, 2, 3, 8, or 9. This is because adult human heights are distributed by a *normal distribution*, not an exponential distribution.

Benford's Law can be used in many practical circumstances to check for fraudulent data. For example, it can be used by the IRS to check tax returns for tax cheaters, by science instructors looking for "invented" lab data, or by economists checking for fraudulent data in economic transactions.

(For more on Benford's Law, see e.g. *Brainteaser Physics* by Göran Grimvall, which inspired the energy use example.)