Épreuve de section européenne

Benford's law

The files on the hard disk of any computer have various sizes (in Kb). If we take a look at them, we can count how many numbers begin with 1, how many begin with 2, how many begin with 3, and so on. You might expect that there would be the same proportion of numbers beginning with each different digit, roughly 1/9, but it is very likely that you will be wrong!

For example, the table below is the result of an experiment on 150,000 files of "My files" folder on Christmas Eve, 2010:

First digit	1	2	3	4	5	6	7	8	9
Number of files	48552	23923	16407	12989	14364	10073	8884	7749	7059
Relative frequency	.324	.16	?	.087	.096	.067	.059	.052	.047

Surprisingly, as for many kinds of data, the distribution of first digits is highly asymmetric, the most common digit being 1 and the least common 9. This fact was discovered in 1881 by the American astronomer Simon Newcomb, by noticing that in logarithm books (used at that time to perform calculations) the earlier pages were much more worn out[1] than the other pages. The phenomenon was rediscovered in 1938 by Franck Benford, a physicist at the General Electric Company, who claimed that the relative frequency of numbers that start with the digit D should be:

$$\log_{10}(D+1) - \log_{10} D \text{ where, for all } x > 0, \log_{10} x = \frac{\ln x}{\ln 10} \text{ is the decimal logarithm of } x.$$

Benford's law is used to track down fraud in various domains; it has been invoked as evidence of fraud in the 2009 Iranian elections. In June 2010, consultants working for political website *Daily Kos* used Benford's law, among other tools, to find serious flaws in the data collected by polling company Research 2000 (R2K). This led to the termination of R2K's contract with Daily Kos...

Adapted from various sources, (Ted Hill's website, Wikipedia, *Plus* magazine).

Questions

1. Why didn't the 2010 Christmas table take into account the digit 0?
2. Explain how Newcomb discovered the "First digit phenomenon".
3. Compute the relative frequency missing in the table.
4. Using Benford's formula, compute the theoretical frequency for each digit. Does the 2010 Christmas computer experiment match the criteria?
5. Should you use Benford's law to choose your lottery numbers? Would you use it with the age of politicians in the French *Assemblée Nationale*? Explain your answers.
6. Describe a simple argument in favor of using Benford's law to trace down fraud.

[1] worn out = in bad shape