Do CNF files satisfy Benford's Law?

Lukas Prokop

18th of August 2016

1 Motivation

Let $(n_1, n_2, \ldots)$ be an arbitrary sequence of numbers. Let $i \in \{0, \ldots, 9\}$ be an arbitrary digit. By the Law of Large Numbers, every digit is equally likely to occur in this sequence of numbers if the data set is large enough. Let $\mathcal{P}_1(i)$ be the occurence probability of a digit i for $i \in \{0, \ldots, 9\}$. Then it holds that

$$\mathcal{P}_1(i) = \frac{1}{10} \qquad \forall i \in \{0, \ldots, 9\}$$

Now consider leading digits (i.e. the digit at most significant position). Intuitively, many people claim that all digits (except zero of course) occur the same number of times. Let $\mathcal{P}_2(i)$ be the corresponding occurence probability. Then

$$\mathcal{P}_2(i) = \frac{1}{9} \qquad \forall i \in \{1, \ldots, 9\}$$

2 Benford's Law

Now consider a large data set containing numbers. Conventionally data from newspapers, logarithmic tables, heights of buildings or people is used. Benford's Law is claimed to hold for any sufficiently collection of numbers.

Contrary to the uniform distribution given with $\mathcal{P}_2$, Benford's Law claimes that the digit 1 occurs in about 30 % of the time. A paper by Theodore P. Hill [3] characterizes Benford's Law as the following conjecture:

> **Conjecture.** *The first digit is* 1 *about* 30 *percent of the time and* 9 *only about* 4.6 *percent of the time.*

Furthermore the paper puts that "that the digits are not equally likely to appear comes as something of a surprise, but to claim an exact law describing their distribution is indeed striking". A more formal definition on Wikipedia [6] claims that digit i should occur with probability close to $\mathcal{P}_3(i) = \log_{10}(1 + 1/i)$.

$$\begin{array}{lll} \mathcal{P}_3(1) = 30.1 & \mathcal{P}_3(2) = 17.61 & \mathcal{P}_3(3) = 12.49 \\ \mathcal{P}_3(4) = 9.69 & \mathcal{P}_3(5) = 7.92 & \mathcal{P}_3(6) = 6.69 \\ \mathcal{P}_3(7) = 5.8 & \mathcal{P}_3(8) = 5.12 & \mathcal{P}_3(9) = 4.58 \end{array}$$

3 Data set

The only reason I had come up with verifying this conjecture was because I had lots of CNF files on my hard disk drive and thought about something fun. As far as literals are integers, I looked for statistics with integers and remembered James Grime's video on Benford's Law [1].

Now consider all public CNF benchmarks between 2008 and 2016. Furthermore consider the dataset by SATlib [2]. The latter constitutes 78.2 % of the data whereas the second-largest repository (SAT competition 2016) contained 8.6 % of the files. Only files with file extension `.cnf` were considered, which is common for CNF files.

I provided zsh scripts to download those files easily [4]. With those scripts it is documented which files are duplicates. Only one copy of duplicates has been considered. Furthermore it is documented that a few individual archives are corrupt and could not be decompressed.

4 Results

Given the data set, we want to verify whether the leading digit is 1 in about 30 % of the time and 9 in about 4.6 % of the time.

In total 90,038,017,631 digits and 21,711,280,112 leading digits in CNF files were found. If we exclude zero, we have 77,632,879,380 digits and 16,465,072,745 leading digits in total. The evaluating software is available online [5].

Table 2 reveals the interesting number to draw our conclusion: Considering a deviation of 1.34 % for digit 1 (expected 30.1, actual 28.76) and 0.12 % for digit 9 (expected 4.58, actual 4.46), we conclude:

Result. *Yes, the CNF files considered satisfy Benford's law.*

References

[1] James Grime. *Benford's Law - How mathematics can detect fraud!* URL: `https://www.youtube.com/watch?v=vIsDjbhbADY`.

[2] hh. *SATLIB - The Satisfiability Library.* URL: `http://www.satlib.org/`.

[3] Theodore P Hill. "The First Digit Phenomenon A century-old observation about an unexpected pattern in many numerical tables applies to the stock market, census statistics and accounting data". In: *American Scientist* 86.4 (1998), pp. 358–363.

[4] prokls. *Download public CNF benchmark files with zsh scripts.* URL: `https://github.com/prokls/cnf-files-download/`.

[5] Lukas Prokop. *Go 001 Verify Benford's Law.* URL: `http://lukas-prokop.at/proj/snippets/go001.html`.

[6] the free encyclopedia Wikipedia. *Benford's Law.* URL: `https://en.wikipedia.org/wiki/Benford%27s_law`.

digit	occurences	percentage	percentage without zero
0	12405138251	13.78 %	—
1	11881104381	13.20 %	15.30 %
2	10126291630	11.25 %	13.04 %
3	9145843706	10.16 %	11.78 %
4	8612466623	9.57 %	11.09 %
5	7996036260	8.88 %	10.30 %
6	7639816582	8.49 %	9.84 %
7	7495610218	8.32 %	9.66 %
8	7366477797	8.18 %	9.49 %
9	7369232183	8.18 %	9.49 %

Table 1: Digit frequency

digit	leading occurence	percentage	percentage without zero
0	5246207367	24.16 %	—
1	4736061175	21.81 %	28.76 %
2	3143734515	14.48 %	19.09 %
3	2232482931	10.28 %	13.56 %
4	1806648285	8.32 %	10.97 %
5	1231553998	5.67 %	7.48 %
6	948825030	4.37 %	5.76 %
7	843834072	3.89 %	5.12 %
8	788172419	3.63 %	4.79 %
9	733760320	3.38 %	4.46 %

Table 2: Leading digit frequency