Physica A 512 (2018) 305–315

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Benford's law and first letter of words

Xiaoyong Yan [a,b], Seong-Gyu Yang [c], Beom Jun Kim [c], Petter Minnhagen [d,*]

[a] *Systems Science Institute, Beijing Jiaotong University, Beijing 100044, China*
[b] *Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, China*
[c] *Department of Physics, Sungkyunkwan University, Suwon 16419, Republic of Korea*
[d] *IceLab, Department of Physics, Umeå University, 901 87 Umeå , Sweden*

HIGHLIGHTS

- Show that the first letters in many English novels follow a universal frequency ladder.
- Show that the universal frequency ladder only depends on the number of letters in the alphabet.
- Point out the similarity to Benford's law and the number of digits.

ARTICLE INFO

Article history:
Received 11 July 2017
Received in revised form 14 June 2018
Available online xxxx

Keywords:
First-Letter Law
Benford's law
Universal frequency ladder
Random group formation
Maximum entropy

ABSTRACT

A universal First-Letter Law (FLL) is derived and described. It predicts the percentages of first letters for words in novels. The FLL is akin to Benford's law (BL) of first digits, which predicts the percentages of first digits in a data collection of numbers. Both are universal in the sense that FLL only depends on the numbers of letters in the alphabet, whereas BL only depends on the number of digits in the base of the number system. The existence of these types of universal laws appears counter-intuitive. Nonetheless both describe data very well. Relations to some earlier works are given. FLL predicts that an English author on the average starts about 16 out of 100 words with the English letter '*t*'. This is corroborated by data, yet an author can freely write anything. Fuller implications and the applicability of FLL remain for the future.

© 2018 Published by Elsevier B.V.

1. Introduction — Benford revisited

Benford's law predicts relative occurrence of first digits for numbers in a data collection [1]. The fact that a data collection of numbers often follows a specific universal distribution of first digits is rather counter-intuitive: it is sometimes unknown to persons involved in creating faked accounts and in such cases the law can be used to uncover such frauds [1]. The law can be expressed as

$$p_i = \log_X\left(\frac{i+1}{i}\right),$$

where X is the base for the number system, $i = 1, 2, \ldots, X-1$ is the first digit of a number and p_i the fraction of numbers which starts with the digit i. For example, the most common first digit is according to this law is 1 and has the relative frequency $p_1 = \log_X 2$. For the usual decimal system with $X = 10$ this gives $p_1 = \log_{10} 2 \approx 0.301$, which means that 30.1% of the numbers in the collection is predicted to start with the digit 1. If the numbers are instead given within the binary system with $X = 2$, then $p_1 = \log_2 2 = 1$, since in the binary system all numbers must start with the digit 1.

[*] Corresponding author.
E-mail address: Petter.Minnhagen@physics.umu.se (P. Minnhagen).

https://doi.org/10.1016/j.physa.2018.08.133
0378-4371/© 2018 Published by Elsevier B.V.

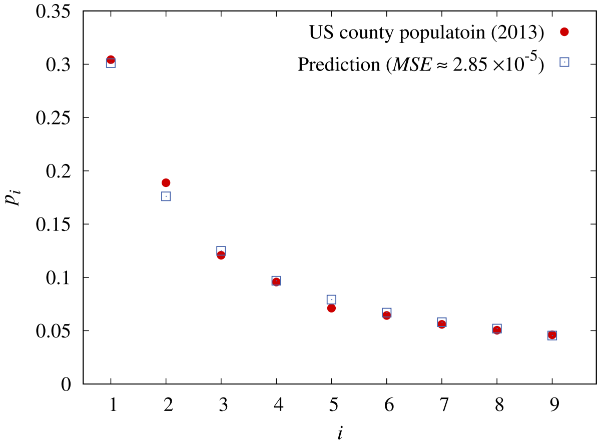

Fig. 1. Distribution of first digits for the number of people belonging to US counties (filled circles, from US population census 2013 [2]) compared to the Benford's law prediction (open squares). MSE (Mean Square Error between the data and the prediction) is 2.85×10^{-5}.

The surprising thing is that the Benford's law to a good approximation is borne out by wide range of data collections from very different contexts [1]. Fig. 1 illustrates the Benford's law in the case of the size distribution of counties in USA. Mean Square Error (MSE) between the data and the Benford's law prediction is small which is about 2.85×10^{-5} for 2013 US county-size distribution. The fact that Benford's law is borne out by a wide range of systems, and that the law carries absolutely no specific information on the system itself, compared with the fact that it has no free parameters, points to a general origin [1].

Benford independently made his discovery around 1938 [3]. However, the law had been discovered earlier by Newcomb around 1880 [4]. Newcomb's discovery allegedly originated from his observation that the logarithm table in a public library used by many, was more worn in the beginning, where people had been looking for logarithm for numbers starting with 1 and less at the end, which gave the logarithms for numbers starting with 9. Thus, the numbers looked up by the users of this logarithmic table were in fact collected into groups labeled by the starting digit. The relation between the various group sizes should, according to Newcomb, be given by Benford's law [4].

Newcomb's observation ties into the present investigation in the following way: Suppose that, instead of looking up logarithms in a table of logarithms, you were looking up words in a dictionary. Suppose you were reading the novel *Moby Dick* by Herman Melville and that you did not know English, so that you had to look up every word in a dictionary into your native language. Furthermore assume that your memory was so bad that you had to look up every word each time occurred in the text. This would mean that you effectively collected words into the first letter groups. These groups would be in different sizes just as the first digit groups for numbers: you would look up more times in the dictionary for certain starting letters than for others. One may then ask if there is some law, akin to Benford's law, which gives the distribution of the group sizes for the first letters.

The First-Letter Law (FLL), derived and discussed in the present paper, is such a law. It can be expressed as

$$p_i = \frac{X - (X-1)\log_X(X-1) - i\log_X i + (i-1)\log_X(i-1)}{X(X-1)\log_X\left(\frac{X}{X-1}\right)}. \tag{1}$$

Here $X(\geq 2)$ is the number of letters in the alphabet, p_i is the ratio of the first letter group i where $i = 1$ is the most frequent first letter and $i = X$ is the least frequent one. According to our First-Letter Law, Eq. (1), the ratio of the most frequent first letter for an English novel is

$$p_1 = \frac{1 + 25\log_{26}\left(\frac{26}{25}\right)}{26 \cdot 25\log_{26}\left(\frac{26}{25}\right)},$$

since the English alphabet has $X = 26$ letters. This gives $p_1 \approx 0.166$. Table 1 gives FLL-prediction for a 26-letter alphabet as the percentage for the occurrence of a first letter of a word in terms of its rank ($i = 1$: the most common first letter, $i = 2$: the second most common, and so on).

Fig. 2 shows the validity of our FLL for the case of the novel *Moby Dick*. The overall agreement is strikingly good and the most frequent first letter has $p_1 = 0.164$ (which is the letter 't'), very close to the predicted ratio 0.166. Table 1 gives all percentages for the first letters in *Moby Dick* and other novels.

The structural similarity between Benford's law (BL) and FLL is that neither of them contains any free parameter: BL only has the number of digits in the numerical base as its input and FLL only the number of letters in the alphabet.

Neither BL nor FLL are always valid: Benford's law has been thoroughly investigated with multitude of examples. Its limitations have been investigated and possible ways of understanding its origin have been proposed and debated [1,5–7].

Download English Version:

https://daneshyari.com/en/article/7374584

Download Persian Version:

https://daneshyari.com/article/7374584

Daneshyari.com