Medical Hypotheses 97 (2016) 102–106

Contents lists available at ScienceDirect

Medical Hypotheses

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

Original article

The "first digit law" – A hypothesis on its possible impact on medicine and development aid

Gregor Pollach *, Frank Brunkhorst, Mwapatsa Mipando, Felix Namboya, Samson Mndolo, Thomas Luiz

University of Malawi (College of Medicine, Department of Anaesthesia and Intensive Care), College of Medicine, BP 360, Chichiri, Blantyre 3, Blantyre, Malawi

ARTICLE INFO

Article history:
Received 23 May 2016
Accepted 24 October 2016

Keywords:
First digit law
Statistics
Benford
Natural data sets
Fraud

ABSTRACT

The "first digit law" or "Benford's law" is a mathematical distribution discovered by Simon Newcomb and Frank Benford. It states, that the probability of the leading number d ($d \in \{1,\ldots,9\}$) in many natural datasets follows: $P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}(1 + 1/d)$.

It was successfully used through tax authorities and "forensic accounting" in order to detect fraud and other irregularities. Benfords law was almost neglected for its use outside financial accounting. The planning for health care systems in developing countries is extremely dependant on good, valid data. Whether you plan the catchment area for the future district hospitals, the number of health posts, the staff establishment for the central hospital or the drug budget in the Ministry. The "first digit law" can be used in medicine, public health, physiology and development aid to unmask questionable data, to discover unexpected challenges, difficulties in the data collection process, loss through corruption and criminal fraud. Our hypothesis suggests, that the "first digit law" is a cost effective tool, which is easy to use for most people in the medical profession, which does not really needs complicated statistical software and can be used on the spot, even in the resource restricted conditions of developing countries. Several preconditions (like the size of the data set and its reach over more than two dimensions) have to be fulfilled, but then Benfords law can be used by any clinician, physiologist, public health specialist or aid consultant without difficulties and without deeper statistical knowledge in the four steps, we suggest in this article. The consequences will be different depending on the level (local regional, national, continental, international) on which you will use the law. All levels will be enabled to get insight into the validity of the data-challenges for the other levels without the help of trained statisticians or accountants. We believe that the "first digit law" is a vastly underestimated and neglected, but extremely useful tool for the identification of unexpected challenges, supervision and control in various parts of medicine and public health for almost all aspects of development aid.

Introduction

In 1881 Simon Newcomb, an American mathematician published a paper called "Note of the Frequency of Use of the Different Digits in Natural Numbers" in the American Journal of Mathematics [1]. This paper was than more or less forgotten and the mathematical law behind it was only rediscovered by Frank Benford (a physicist at General Electric's) 1938 in his article "The law of anomalous numbers" in the proceedings of the American Philosophical Society [2]. Since than we usually call the mathematics behind it "Benford's law" or the "first digit law". The law states, that in many naturally occurring datasets (baseball statistics, bank accounts, the Fibonacci numbers, the area of rivers in a country, physical or physiological constants, or the amount of taxes payed in the UK etc. [2–4]) the nominal value of the first digit of all the single data follows an somehow counterintuitive, but extremely interesting rule. When you control for all numbers occurring in a newspaper and than you count how often you will find the "1", or the "2" – "9" as first digit (the first digit in 16789 is "1") you will realize that they do not occur with a frequency of 11.1% (100 divided by 9 [1–9]) as many people would expect, but by a distribution far from it.

Benford states, that they will occur with a probability for the leading number d ($d \in \{1,\ldots,9\}$), following the "first digit law", as: $P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}(1 + 1/d)$ [5].

The "1" therefore will occur as the leading number in around 30.1% of the cases; the number "2" in around 17.6%, the number "3" in 12.5%, the number "4" in 9.7%, the number "5" in 7.9%, the

* Corresponding author.
E-mail addresses: gpollach@medcol.mw, gipi.bc62@yahoo.de (G. Pollach).

http://dx.doi.org/10.1016/j.mehy.2016.10.021
0306-9877/

number "6" in 6.7%, the number "7" in 5.8%, the number "8" in 5.1% and the number "9" in 4.6% [5].

Hypothesis

Our hypotheses states that we have found a mathematical tool for quantitative data management and quality control which was not widely used in medicine, physiology, public health and development aid until today. Our hypothesis is that through the counterintuitive distribution of the first number in the Newcomb/Benford law we can use the distribution of this first digit in many medical datasets to discover unexpected challenges, difficulties in the data collection process, loss through corruption or to unmask questionable data or deliberate fraud.

Through our "four step approach" we provide a practical mechanism to implement it on each level; might that be locally, regional, national, continental or internationally.

Initially it would be prudent to narrow the field of applications (e.g. for the clinician) to his direct surroundings in his institution. This could be in a hospital the budget the director need for the travelling to the Ministry of Health's Headquarters compared to the money allocated to mother and child care or for the doctor in the district the question, whether relatively to Ebola (not existing at all in most countries) related issues the correct amount was spend on Malaria treatment.

The "first-digit law" shows us a new, quick, always available, cheap and easy to understand view on data and additionally provides us with a "field laboratory" like way to scrutinizing them.

The basis of the hypotheses/deduction

Despite some suggestions to do so (Varian 1972 in "The American Statistician"[7]) there were few papers during the last decades [8] trying to use the "first digit law" in real life situations (might it be for the rich or the resource poor countries). The notable exception have been our tax authorities, probably influenced by Nigrinis paper "Taxpayers Compliance Application of Benford's Law [9]". It was additionally used especially in the field which calls itself "forensic auditing". Forensic auditing used Benford's law to determine whether they deal with natural datasets or with datasets after they were "treated" fraudulently. Another exception were Judge et al. [10] trying to use Benford law for economic data or Pollach [11] for pediatrics. Not the least success of forensic accounting was, albeit retrospectively, the discovery that the Greek government did not present their real economical data before they entered the European Union [12,13].

On this basis we suggest that the "first digit law" should be used to scrutinize data in clinical medicine, public health, research and project funding.

We suggest as a practical example for the importance of the proposed field approach an obstetrician in an African country who wants to know, without really deep statistical knowledge and without spare time, whether the data on maternal mortality in his country are correct or whether they suffer from major misbalances.

He could e.g. compare all MMR data from the different health centers through the first digit law. MMR is one of the most important "numbers" in general obstetrics and the sustainable development for vast regions of the world. It is very important for public health and provides a well acknowledged basis for important political decisions.

The obstetrician in our example might discover through Benfords law challenges in the distribution of donated money to the various districts. So he received the information, that some districts might receive too much help, which is lacking in the other districts maternities or that the help for maternal care in the district is not in an equilibrium with the funds for anaesthesia or peds. Funds are scarce and need a distribution following the needs. On the other hand it is in daily work in the tropics a great relieve to find budgets in ones own hospital or districts which hold after scrutiny. Moreover he might find problems in the data collection process itself or even fraud in the funding agencies – all of which have important influences on his future work.

Footnote: (The obstetrician was chosen, because obstetrics is one of the most important medical profession in developing countries. Moreover they receive a high percentage of the international development aid –and indeed the law can help to unmask such a fraud).

Evaluation of the hypotheses

Even when we argued above that the distribution following Benfords law is somehow counterintuitive we do think that we can explain them in a way that they can be considered intuitively.

An example

The Ministry of Health in a resource poor country allocates for several years on a special account (in order to guarantee maintenance sustainability) for each district hospital the amount of "x" US $ (let us say 10,000 $). For the sake of easy mathematical operations let us assume that the interest rate is 100%. This means that the account should present 20,000 $ at the end of the first year, 40,000 at the end of the second and 80,000 at the end of the third year. This means too that a monthly budget scrutiny which would look at the first digit on the account will see during the whole first year a "1" (from 10,000 at the first day until 19,999 – only the last day of the year will show a "2" for 20,000 $). During the second year we start with a "2" until we reach during the year 29,999 and than the "3" for 30,000 and will see in the end of the year a "4" for 40,000. In the third year we start with a "4", pass the "5"0000, "6"0000 and "7"0000 in order to end up with an "8" for 80,000. During the next years the same is repeated with the first digits for the hundreds of thousands and the millions. Lets assume the MoH will really leave the money for ten years on the account and then employs an auditor who controls the amount of money at 1000 points of time during the 10 years it is now understandable that he will find much more often the "1" as first digit than the "2"- and this one more often than the "3" and so on. Here we can see that this distribution is a logarithmic one. The digit "1" appears in 30.1%. This is the result of $\log_{10}(2) - \log_{10}(1)$. The digit "2" appears in 17.6% of cases which correspond to $\log_{10}(3) - \log_{10}(1) = 0.176$ (Figs. 1 and 2).

Here we come back to the initial problem of Simon Newcomb. In 1881 when logarithmic tables were widely used, he discovered that the first pages of his personal copy were dirtier than the last ones – meaning that the pages beginning with the digit one were being looked up much more frequently than the others; a discovery which then led to the "first digit law" [1].

The planning for health care systems in developing countries is extremely dependant on good, valid data. This is true, whether you plan the catchment area for the district hospital you need in your region, the number of health posts, the availability of health surveillance assistants, the size of specialized units in the central hospitals, the amount of drugs you have to put into central medical stores, the budget planning for physical assets management or the planning for the intake of your College of Medicine. These data are usually hard to find and even much harder to interpret. In fact it is for the non-specialized statistician or accountant in the tropics usually impossible to interpret them correctly.