International Journal of Applied Mathematics

Volume 33 No. 4 2020, 685-695
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
doi: http://dx.doi.org/10.12732/ijam.v33i4.12

LIMITS OF BENFORD'S LAW IN EXPERIMENTAL FIELD

Stéphane Blondeau Da Silva

13 François Chnieux Street
Limoges 87000, FRANCE

Abstract: Two of the main flaws of Benford's law will be discussed in this article: (i) the first one, which leads the observer to consider an experimental dataset as the result of a single random variable rather than several, makes this law so mysterious; (ii) the second one is that Benford's probabilities have long been considered as perfect values: this is obviously not the case.

AMS Subject Classification: 60E05
Key Words: Benford's law; leading digit; experimental data

1. Introduction

Benford's Law, also called NewcombBenford's law, is really remarkable: it states that the first digit d, $d \in [\![1, 9]\!]$, of numbers in many naturally occurring collections of data does not follow a discrete uniform distribution, as might be thought, but a logarithmic distribution (see the recent books of Miller [41] and Berger and Hill [10]). Evoked by the astronomer Newcomb in 1881 ([43]), this law was definitively brought to light by the physicist Benford in 1938 ([8]). He proposed the following distribution: the probability for d to be the first digit of a number is equal to $\log(d+1) - \log(d)$, i.e. $\log(1 + \frac{1}{d})$. Benford tested it over a data set from 20 different domains (surface areas of rivers, sizes of American populations, physical constants, molecular weights, entries from a mathematical handbook, numbers contained in an issue of Reader's Digest, the street addresses of the first persons listed in American Men of Science, death rates, etc.).

Received: April 28, 2020

Numerous empirical data, as economic data ([55, 16]), social data ([31, 42, 28]), demographic data ([48, 39]), physical data ([36, 13, 47, 52, 18, 3, 1]), biological data ([17, 26, 37]), data in psychology ([20, 6, 15]) or internet data ([4]), for instance, conform to Benford's law ([34, 46]); to such an extent that this law was used to detect possible frauds in lists of socio-economic data ([59, 45, 22, 53, 58, 51, 14, 38]) or in scientific publications ([2]).

Nevertheless many voices were quick to challenge this overly consensual message... First of all, many empirical datasets are known to fully disobey Benford's law ([50, 32, 57, 54, 7, 19]). In addition to that, this law often appeared to be a good approximation of the reality, but no more than an approximation ([54, 53, 19, 27, 29]). Goodman, for example, in [29], discussed the necessity of introducing an error term. Even the 20 different domains, tested by Benford (in [8]), displayed large fluctuations around theoretical values.

Thus, this law does not seem to be a universal law but it appears in certain specific contexts ([54, 40, 21, 12, 5]). And the groundings and applicability of the Benford law remain highly obscure [60, 9].

In this article we will attempt to counter two of the main errors concerning Benford's law: (i) the first one is to consider that data derives from a single random variable, belief making Benford's law particularly mysterious; (ii) the second is related to probability values themselves which are questionable.

2. Counter-intuitive yes but not mysterious at all

Curious, surprising, mysterious, a significant number of such qualifiers have, since its appearance, characterized Benford's law and still characterizes it; the title of the article published by Benford himself is by the way: *The law of anomalous numbers* ([8])! But, even if it is really counterintuitive, it is all but mysterious.

One of the main error that is made when observing certain experimental datasets is to believe that the data is derived from a single random variable; with a single random variable, the frequencies of each of the nine possible first digits could be quite balanced (especially if the random variable is following a uniform law or even if it covers many orders of magnitude).

But in numerous cases, this assumption, which is intuitively made, is not valid, even if, at first glance, only one random variable seems to underlie the data. Let's take a few examples; among the domains chosen by Benford himself ([8]), some are undeniably the result of many independent random variables: physical constants, entries from a mathematical handbook, numbers contained

in an issue of Reader's Digest, etc. Others, in particular natural quantities, depend on such a large number of parameters that data can be considered as the result of multiple (at best non-correlated) random variables: surface areas of rivers, sizes of american populations, etc.

Emerging from multiple random variables rather than a single one is crucial to the data: conditions for Benford's law to reach its full potential are in this way gathered; indeed, by choosing numbers according to randomly selected distributions, the resulting list will obey Benford's Law ([33, 34]). Considering mixtures of uniform distributions, an explanation for the appearance of Benford's Law in everyday-life numbers has already been advanced ([24, 30, 35, 11]). In an article entitled "sufficient conditions for Benford's law" ([5]), Balanzario and Sánchez-Ortiz point out that numerous independant random variables make Benford's law relevant.

Let us now focus more specifically on the case of addresses first mentioned by Benford in [8], but also mentioned in [35, 11]. In that case, the distribution of empirical data is naturally modelled by a sample of uniform distributions: drawing a random address number in a given street gives, indeed, gives rise to a random variable uniformly distributed.

In the simplest of cases, more specifically studied in [11], the choice of the street itself can follow an uniform distribution. Thus, two steps are involved in the selection of the street number: (i) the drawing of a street, from a uniform distribution on $\{1, ..., H\}$, the selected number h being the highest number in the considered street (the number H represents the maximum number that street numbers can reach), (ii) the drawing of the street number in that selected street, it follows a uniform distribution on $\{1, ..., h\}$. There are thus two successive draws following discrete uniform laws according to this model (see Figure 1 and [11]).

Let S be the random variable that maps any draw pair to the highest numbers of the selected streets (the pair first component) and N be the random variable that maps them to the chosen street numbers (the pair second component).

The distribution of the street numbers is as follows:

$$\begin{aligned}\forall k \in \{1, ..., H\},\ \mathrm{P}(\{N = k\}) &= \sum_{h=1}^{H} \mathrm{P}(\{N = k\} \cap \{S = h\}) \\ &= \sum_{h=1}^{H} \mathrm{P}(\{S = h\}) \times \mathrm{P}_{\{S=h\}}(\{N = k\})\end{aligned}$$

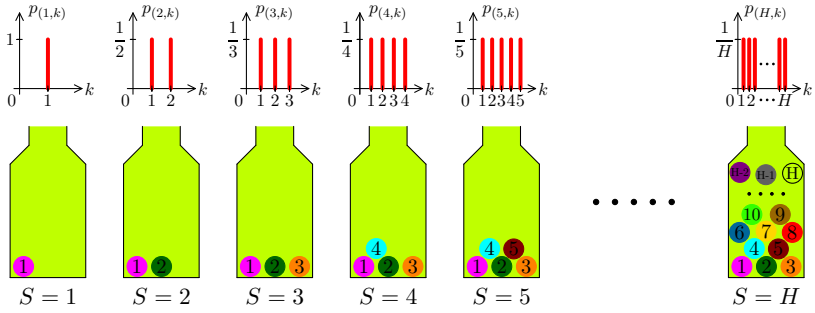

Figure 1: Representation of Blondeau Da Silva's model with two successive draws. The first draw consists in selecting an urn (i.e. a street), the second one consists in selecting a ball in this urn (i.e. a street number). For $h \in \{1, ..., H\}$, $p_{(h,k)} = \mathrm{P}_{\{S=h\}}(\{N = k\})$ are the probabilities of the second draw (in the case where $S = h$), the random variable N following a discrete uniform distribution on the integers 1, ..., h.

$$= \frac{1}{H} \sum_{h=k}^{H} \frac{1}{h}.$$

Figure 2 shows the shape of the distribution (for $H = 739$).

It has been observed that other factors favour the relevance of Benford's law; as a matter of fact, it tends to apply more accurately to data that spans several orders of magnitude ([23, 56]). For its part, Formann showed that certain random variables with long right-tailed distributions are compatible with Benford's law ([25]): it is interesting to note that the distribution resulting from the model studied above is long right-tailed (see Figure 2)!

Considering the distribution of Figure 2, probabilities of drawing each digit can be determined. They are gathered in Table 1.

Probabilities determined by the model in [11] seem to be close to those defined by Benford [8] (see Table 1). Thus, this model is relevant and rather accurate.

It notably unveils the mystery surrounding the existence of predominances in experimental data sets: the predominance of 1 over 2, 2 over 3, 3 over 4 etc. It is simply due to the fact that in the lexicographical order, 1 appears before 2, 2 before 3, 3 before 4 etc.: if we consider the street whose address numbers range from 1 to 200, the proportion of 1 as first digit is of course very high: 0.555! Benford's law loses some of its mystery. Counter-intuitive yes,

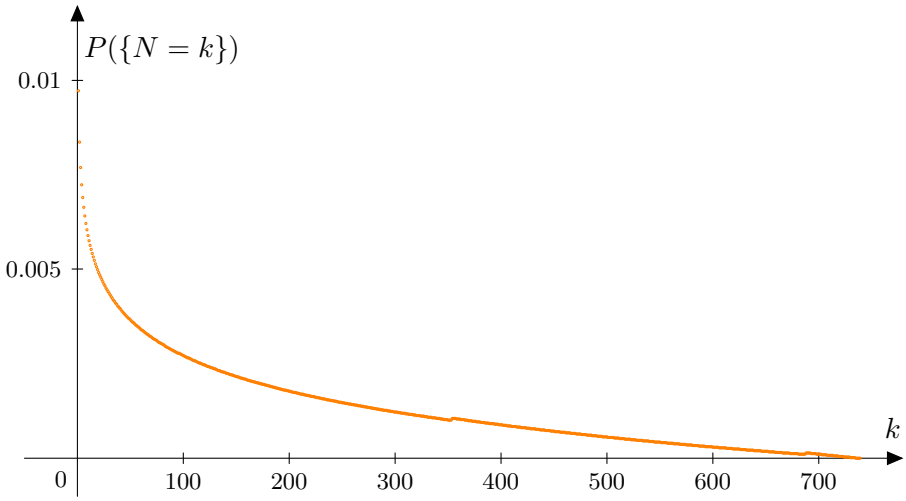

Figure 2: The distribution of address numbers when $H = 739$.

mysterious no!

3. Credence have been given to theoretical values in the analysis of experimental datasets

We can see that probabilities of the proposed model differ from Benford's ones, even if they are close (see Table 1).

In [11], Blondeau Da Silva showed that, the proportion of each d as leading digit, $d \in [\![0, 9]\!]$, structurally fluctuates: they do not tend to Benford's famous values.

Benford's probabilities through their simple expression and therefore their easy use, has gained great prominence. But its importance has been largely overestimated. In the above model, these probabilities are never reached. The literature, as seen in the introduction, is rich in examples in which Benford's values are only approximations of the frequencies of first digits in experimental data sets.

This is the second major misunderstanding about Benford: a sometimes blind faith in values that are too simple.

d	First digit probability	$\log(1+\frac{1}{d})$
1	0.283	0.301
2	0.203	0.176
3	0.151	0.125
4	0.113	0.097
5	0.083	0.079
6	0.058	0.067
7	0.039	0.058
8	0.036	0.051
9	0.034	0.046

Table 1: Values of d, probabilities that the first digit of the address number is d the maximum adress number value being 739 ([11]) and Benford's law probabilities ([8]), these values being rounded to the nearest thousandth.

Conclusion

Finally, Benford's law does not seem as mysterious as expected, the predominance of small figures as first digit being due to the lexicographical order. Considering an experimental dataset as the realization of several random variables rather than a single one allows us to understand this observation. We must also be careful with Benford's values, which sometimes give a correct idea of the distribution of digits, but nothing more. Numerous datasets do not or only slightly follow this law. The temptation to reduce digits frequencies in empirical data to excessively simple values is reassuring, but it is a pitfall that has been encountered many times in science.

The Pareto principle on experimental data is no exception: it states that, for many events, roughly 80% of the effects come from 20% of the causes. The 80/20 rule is approximately followed by a power law distribution, the Pareto distribution, and many natural phenomena have been shown empirically to exhibit such a distribution ([44]). But as with Benford's law, many reservations are expressed. The Pareto principle is an empirical principle, therefore imprecise. Economist and sociologist Thomas Piketty recalls this mathematical evidence in the case of its use concerning the distribution of wealth: "Even today, some people still imagine, following Pareto, that the distribution of wealth would be characterized by an implacable stability, the consequence of an almost divine law. In truth, nothing could be further from the truth: when inequalities

are studied from a historical perspective, what is important and what needs to be explained is not the slight stability, but rather the considerable changes" [49].

References

[1] T. Alexopoulos, S. Leontsinis, Benford's law in astronomy, *Journal of Astrophysics and Astronomy*, **35**, No 4 (2014), 639-648.

[2] A. D. Alves, H. H. Yanasse, N. Y. Soma, Benford's law and articles of scientific journals: comparison of JCR and Scopus data, *Scientometrics*, **98** (2014), 173-184.

[3] J. Aron, Crime-fighting maths law confirms planetary riches, *New Scientist*, **220** (2013), 12-13.

[4] L. Arshadi, A. H. Jahangir, Benford's law behavior of internet traffic, *Journal of Network and Computer Applications*, **40** (2014), 194-205.

[5] E. P. Balanzario, J. Snchez-Ortiz, Sufficient conditions for Benford's law, *Statistics ans Probability Letters*, **80**, No 23-24 (2010), 1713-1719.

[6] G. Beeli, M. Esslen, L. Jncke, Frequency correlates in grapheme-color synæsthesia, *Psychological Science*, **18** (2007), 788-792.

[7] T. W. Beer, Terminal digit preference: beware of Benford's law, *Journal of Clinical Pathology*, **62**, No 2 (2009), 192.

[8] F. Benford, The law of anomalous numbers, *Proceedings of the American Philosophical Society*, **78** (1938), 127-131.

[9] A. Berger, T. P. Hill, Benford's law strikes back: No simple explanation in sight for mathematical gem, *The Mathematical Intelligencer*, **33**, No 1 (2011), 85-91.

[10] A. Berger, T. P. Hill, *An Introduction to Benford's Law*, Princeton University Press, Princeton (2015).

[11] S. Blondeau Da Silva, Benford or not Benford: a systematic but not always well-founded use of an elegant law in experimental fields, *Communications in Mathematics and Statistics*, **8** (2020), 167-201.

[12] A. Bonache, J. Maurice, K. Moris, Dtection de fraudes et loi de Benford : Quelques risques associes, *Revue Franaise de Comptabilit*, **431** (2010), 24-27.

[13] J. Burke, E. Kincanon, Benford's law and physical constants: the distribution of initial digits, *American Journal of Physics*, **59** (1991), 952.

[14] R. S. Cella, E. Zanolla, Benford's law and transparency: an analysis of municipal expenditure, *Brazilian Business Review*, **15**, No 4 (2018), 331-347.

[15] M. C. Chou, Q. Kong, C. P. Teo, Z. Wang, H. Zheng, Benford's law and number selection in fixed-odds numbers game, *Journal of Gambling Studies*, **25** (2009), 503-521.

[16] P. Clippe, M. Ausloos, Benford's law and theil transform of financial data, *Physica A: Statistical Mechanics and its Applications*, **391**, No 4 (2012), 6556-6567.

[17] E. Costasa, V. Lopez-Rodasa, F. Torob, A. Flores-Moya, The number of cells in colonies of the cyanobacterium microcystis aeruginosa satisfies Benford's law, *Aquatic Botany*, **89**, No 3 (2008), 341-343.

[18] S. Cournane, N. Sheehy, J. Cooke, The novel application of Benford's second order analysis for monitoring radiation output in interventional radiology, *Physica Medica*, **30** (2014), 413-418.

[19] J. Deckert, M. Myagkov, P. Ordeshook, Benford's law and the detection of election fraud, *Political Analysis*, **19** (2011), 245-268.

[20] S. Dehaene, J. Mehler, Cross-linguistic regularities in the frequency of number words, *Cognition*, **43** (1992), 1-29.

[21] A. Diekmann, B. Jann, Benford's law and fraud detection: Facts and legends, *German Economic Review*, **11**, No 3 (2010), 397-401.

[22] C. Durtschi, W. Hillison, C. Pacini, The effective use of Benford's law to assist in detecting fraud in accounting data, *Journal of Forensic Accounting*, **V** (2004), 17-34.

[23] R. M. Fewster, A simple explanation of Benford's law, *The American Statistician*, **63**, No 1 (2009), 26-32.

[24] B. J. Flehinger, On the probability that a random integer has initial digit a, *American Mathematical Monthly*, **73**, No 10 (1966), 1056-1061.

[25] A. K. Formann, The Newcomb-Benford law in its relation to some common distributions, *Plos One*, **5** (2010).

[26] J. L. Friar, T. Goldman, J. Prez-Mercader, Genome sizes and the Benford distribution, *Plos One*, **7**, No 5 (2012).

[27] N. Gauvrit, J.-P. Delahaye, Scatter and Regularity Imply Benford's Law... and More, *World Scientific* (2011), 53-69.

[28] J. Golbeck, Benford's law applies to online social networks, *Plos One*, **10**, No 8 (2015).

[29] W. Goodman, The promises and pitfalls of Benford's law, *Significance*, **13**, No 3 (2016), 38-41.

[30] A. Herzel, Sulla distribuzione delle cifre iniziali dei numeri statistic, In: *Atti XV e XVI Riunione Sci., Roma* (1956), 205-228.

[31] M. J. Hickman, S. K. Rice, Digital analysis of crime statistics: does crime conform to Benford's law?, *Journal of Quantitative Criminology*, **26** (2010), 333-349.

[32] T. P. Hill, Random-number guessing and the first digit phenomenon, *Psychological Reports*, **62**, No 3 (1988), 967-971.

[33] T. P. Hill, A statistical derivation of the Significant-Digit law, *Statistical Science*, **10**, No 4 (1995), 354-363.

[34] T. 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, *American Scientist*, **86**, No 4 (1998), 358-363.

[35] E. Janvresse, T. De La Rue, From uniform distributions to Benford's law, *Journal of Applied Probability*, **41**, No 4 (2004), 1203-1210.

[36] D. Knuth, *The Art of Computer Programming 2*, Addison-Wesley, New-York (1969).

[37] M. Kreuzer, D. Jordan, B. Antkowiak, B. Drexler, E. F. Kochs, G. Schneider, Brain electrical activity obeys Benford's law, *Anesthesia & Analgesia*, **118** (2014), 183-191.

[38] L. Lacasa, J. Fernandez-Gracia, Election forensics: Quantitative methods for electoral fraud detection, *Forensic Science International*, **294** (2019), e19-e22.

[39] L. Leemis, B. Schmeiser, D. Evans, Survival distributions satisfying Benford's law, *The American Statistician*, **54**, No 4 (2000), 236-241.

[40] T. Lolbert, On the non-existence of a general Benford's law, *Mathematical Social Sciences*, **55** (2008), 103-106.

[41] S. J. Miller, *Benford's Law: Theory and Applications*, Princeton University Press, Princeton (2015).

[42] T. A. Mir, The Benford law behavior of the religious activity data, *Physica A: Statistical Mechanics and its Applications*, **408** (2014), 1-9.

[43] R. Newcomb, Note on the frequency of use of the different digits in natural numbers, *American Journal of Mathematics*, **4** (1881), 39-40.

[44] R. E. J. Newman, Power laws, Pareto distributions and Zipf's law, *Contemporary Physics*, **46**, No 5 (2005), 323-351.

[45] M. Nigrini, I've got your number: How a mathematical phenomenon can help CPAs uncover fraud and other irregularities, *Journal of Accountancy*, (1999).

[46] M. Nigrini, *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*, Vol. 586, John Wiley & Sons, (2012).

[47] M. Nigrini, S. Miller, Benford's law applied to hydrology data-results and relevance to other geophysical data, *Mathematical Geology*, **39**, No 5 (2007), 469-490.

[48] M. Nigrini, W. Wood, Assessing the integrity of tabulated demographic data, *Preprint* (1995).

[49] T. Piketty, *Le Capital au XXIe sicle*, Vol. 585, Seuil (2013).

[50] R. A. Raimi, The first digit problem, *American Mathematical Monthly*, **83**, No 7 (1976), 521-538.

[51] B. Rauch, M. Gttsche, G. Brlher, S. Engel, Fact and fiction in EU-governmental economic data, *German Economic Review*, **12**, No 3 (2011), 243-255.

[52] M. Sambridge, H. Tkalcic, P. Arroucau, Benford's law of first digits: From mathematical curiosity to change detector, *Asia Pacific Mathematics Newsletter*, **1** (2011), 1-6.

[53] A. Saville, Using Benford's law to detect data error and fraud: An examination of companies listed on the Johannesburg Stock Exchange, *South African Journal of Economic and Management Sciences*, **9**, No 3 (2006), 341-354.

[54] P. D. Scott, M. Fasli, Benford's Law: An empirical investigation and a novel explanation, In: *CSM Technical Report 349*, University of Essex (2001).

[55] T. Sehity, E. Hoelz, E. Kirchler, Price developments after a nominal shock: Benford's law and psychological pricing after the euro introduction, *International Journal of Research in Marketing*, **22**, No 4 (2005), 471-480.

[56] S. W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing*, Ch. 34: Explaining Benford's law (2012).

[57] C. Tolle, J. Budzien, R. Laviolette, Do dynamical systems follow Benford's law?, *Chaos*, **10**, No 2 (2000).

[58] K. Tdter, Benford's law as an indicator of fraud in economics, *German Economic Review*, **10** (2009), 339-351.

[59] H. Varian, Benford's law, *The American Statistician*, Letters to the Editor, **26**, No 3 (1972), 62-65.

[60] G. Whyman, E. Shulzinger, E. Bormashenko, Intuitive considerations clarifying the origin and applicability of the Benford law, *Results in Physics*, **6** (2016), 3-6.