Benford's Law of First Digits: From Mathematical Curiosity to Change Detector

Malcolm Sambridge, Hrvoje Tkalčić and Pierre Arroucau

More than 100 years ago it was predicted that the distribution of first digits of real world observations would not be uniform, but instead follow a trend where measurements with lower first digit (1,2,...) occur more frequently than those with higher first digits (...,8,9). This result has long been known by mathematicians but regarded as mere mathematical curiosity. In the physical sciences awareness of Benford's law, as it became known, has been slow to spread. Recently the list of phenomena which follow the predictions of the law has expanded, and physical scientists have begun to find new ways to put it to practical use.

Major scientific discoveries have often resulted from the chance recognition of a pattern or trend in observations. In 1854 J. Snow noticed how cholera patients had all been drinking from the same water pump in London [1]. The recognition of a pattern in data led to the discovery that cholera spread through contaminated drinking water, even though bacteria and viruses were unknown at the time [2]. That breakthrough later led Louis Pasteur to formulate the theory of germs which helped lay the foundations of modern microbiology. Another example is the discovery in 1936 of Earth's inner core by Inge Lehmann [3]. Lehmann noticed something anomalous in the seismic recordings of distant earthquakes. Energy in the form of two new seismic phases were observed at the surface in places where there should only be energy of PKP waves propagating through the liquid outer core. This turned out to be the first observation of new seismic phases that could only have been caused by the presence of a solid inner core. In both cases unexpected patterns seen in observations led to major discoveries.

Today observations are collected at rates never before seen, and scientists are constantly seeking new automated ways to detect subtle signals and extract information from massive data streams. Examples include experiments to unravel the basic forces shaping the universe, e.g. the search for the elusive Higgs particle in the Large Hadron Collider [4]; detection of gravity waves [5]; discovery of new drugs [6] and analysis of the human genome [7]. Recent work by geophysicists [8] has suggested that an intriguing pattern in data, first proposed more than 100 years ago, may provide a new way to detect change in physical phenomena. The pattern in question is known as the first digit, or Benford's law, which itself has been discovered, forgotten and rediscovered over the past century.

In this article we briefly introduce the phenomenon, provide some theoretical insight, outline recent developments and conclude with the suggestion that analysis of digits may provide a novel way of detecting subtle change in data trends across the physical sciences.

1. A Brief History of the First Digit Phenomenon

In the 19th century the astronomer Newcomb [9] first noticed that library books of logarithms were more thumbed in the earlier pages than the latter. He explained how this could arise if the frequency of first digits themselves were not uniform in real world observations but rather followed the rule

$$P_D = \log_{10}\left(1 + \frac{1}{D}\right) \qquad (1)$$

where P_D is the probability of first (nonzero) digit D occurring ($D = 1, \ldots, 9$). For example, the real numbers 123.0 and 0.016 both have $D = 1$, and the digit law suggests that numbers beginning with a 1 will occur about 30% of the time in nature, while those with a first digit of 2 will occur about 17% of the time, and so on down to first digits of 9 occurring about 4% of the time. This decreasing trend of probabilities with digit is represented pictorially in Fig. 1, together with some modern data sets that appear to follow it. The implications

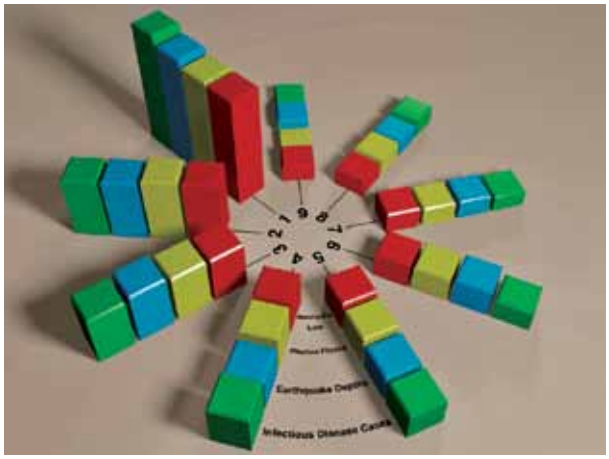

Fig. 1. Benford's law predictions according to (1) for distributions of 1st digits compared to the three data sets, (i) photon fluxes for 1452 bright objects identified by the Fermi space telescope, (ii) 248,915 globally distributed earthquakes in the period 1989–2009, and (iii) 987 reports of infectious disease numbers to World Health Organisation in 2007. Data from [8]. The 1st digit distributions from a wide variety of data sets appear to fit the predictions of the 1st digit law well. [Figure courtesy of Rhys Hawkins, ANU Visualisation laboratory.]

of the digit rule are significant as not only is the distribution not uniform, implying that digit frequencies are not independent, but it must also hold irrespective of the units of the data and their source. Hence a universal property of real world measurements. The result was rediscovered in 1938 by an engineer called Benford [10]. Benford also extended the law to arbitrary base, B, and to multiple digits, N. In this case (1) becomes

$$P_D = \log_B\left(1 + \frac{1}{D}\right) \qquad (2)$$

where the number of possibilities for D depends on both B and N and we have $B^{N-1} \leq D \leq B^N - 1$. For example, for two digits $N = 2$ and there are 90 possibilities for D, i.e. $D = 10, 11, \ldots, 99$. As the number of digits increases the probability distribution becomes flatter and more uniform.

In his original paper Benford showed that 20,229 real numbers drawn from 20 sources all approximately followed the same first digit rule. These included populations of cities, financial data and American baseball league averages. Benford's results were well known in mathematical circles and despite a waning of interest his name became associated with the law. Thirty years later the same first digit distribution was noticed in numbers encountered in computer registries [11]. This led to the suggestion that advanced knowledge of the digit frequency encountered by computers might be used to optimise their design. It has also been suggested that Benford's law may provide a novel way of testing realism in mathematical models of physical processes [12]. If quantities associated with those processes are known to satisfy BL then computer simulations of them should do also. More recently BL has been shown to hold in stock prices [13] and some election results [14].

2. Theoretical Insight

Theoretical insight into the origin and reasons for Benford's law was provided by [15–17, 12]. It is known that BL is the only probability distribution which is both scale and base invariant, properties which such a rule must have to be universally applicable. By scale invariance it follows that if first digits of the variable x follow (1) then so will the first digits of the rescaled variable λx, for any value of λ. Since the Benford distribution is the only one with this property the converse is also true, i.e. if the first digits of x do not follow (1) then no rescaling will make them do so. Scale invariance can be used as a way to measure fit to the law of any infinite sequence of numbers [18]. The Fibbonacci sequence is a well known example, as are many geometric series $x_n = ar^n$ and dynamical systems of the form $x_{n+1} = x_n^2 + 1$ [19].

A second mathematical result is that even if individual distributions of real variables do not follow BL, random samples from those distributions will tend to follow BL, the so called *Random samples from Random distributions* theorem [17].

Over the years there have been a number of mathematical and statistical explanations put forward for Benford's law. One of the earliest was by Feller in his classic 1971 statistical textbook [20], which was later challenged [21]. Recently Fewster [22] has put forward a particularly appealing one for the case of real valued quantities. In that study an experiment is carried out where the probability distribution with the worse possible fit to Benford's law is solved for a given dynamic range of the deviates and smoothness of the PDF. Numerical results show that it becomes increasingly difficult for deviates to fail a goodness of fit test with respect to fitting Benford as smoothness of the PDF increases [22]. At the same time other mathematicians have put forward arguments that smoothness of a PDF alone does not guarantee adherence to Benford's law [21]. For integer se-

quences the origin of Benford's law is less well developed and indeed there appears to be no single explanation that covers all cases [21].

3. Finite Range Real Numbers

One aspect that appears to have received little attention is the influence of the finite range of the data on adherence to Benford's law. It is known that Benford's law of digits will result if the random deviates have a log-uniform modulo 1 distribution. For real valued random deviates that span a single decade, i.e. $1 < x < 10$, and have PDF $P(x)$ we have,

$$P_D = \frac{\int_D^{D+1} P(x)dx}{\int_1^{10} P(x)dx}. \tag{3}$$

If the PDF of the random deviates has a log-uniform distribution or equivalently $P(x) \propto 1/x$ we have

$$P_D = \frac{[\ln x]_D^{D+1}}{[\ln x]_1^{10}} = \frac{\ln\left(1+\frac{1}{D}\right)}{\ln 10} \tag{4}$$

which reduces to (1) and hence Benford's law is recovered. For the case where the deviates span multiple decades, i.e. $10^\alpha \le x \le 10^\beta$ and $\beta > \alpha + 1$ the above integral becomes

$$P_D = \frac{1}{C_{\alpha,\beta}} \sum_{i=1}^{\beta-\alpha} \int_{D\times 10^{\alpha+i-1}}^{(D+1)10^{\alpha+i-1}} \frac{1}{x}\,dx \tag{5}$$

where the normalising constant, $C_{\alpha,\beta}$, is obtained by integrating the probability density over the whole range

$$C_{\alpha,\beta} = \int_{10^\alpha}^{10^\beta} \frac{1}{x}\,dx = (\beta-\alpha)\ln 10. \tag{6}$$

Evaluating these integrals and combining shows that once again P_D reduces to (1) and hence Benford's law is recovered. Therefore the probability of occurrence of each digit is unchanged when the data range extends over an integer number of decades. The situation changes however for the most general case of arbitrary limits, $a \times 10^\alpha \le x \le b \times 10^\beta$. In this case the integral range can be separated into three contributions, the first for the lower end of the range $a\times 10^\alpha \le x \le 10^{\alpha+1}$, the second for the decadal range, $10^{\alpha+1} \le x \le 10^\beta$, and the third for the upper non-decadal part $10^\beta \le x \le b \times 10^\beta$. Evaluating each of these gives the generalisation of Benford's law to arbitrary range log-uniform random deviates

$$P_D = \frac{1}{\lambda_c}\left[(\beta-\alpha-1)\log_{10}\left(1+\frac{1}{D}\right) + \lambda_a + \lambda_b\right] \tag{7}$$

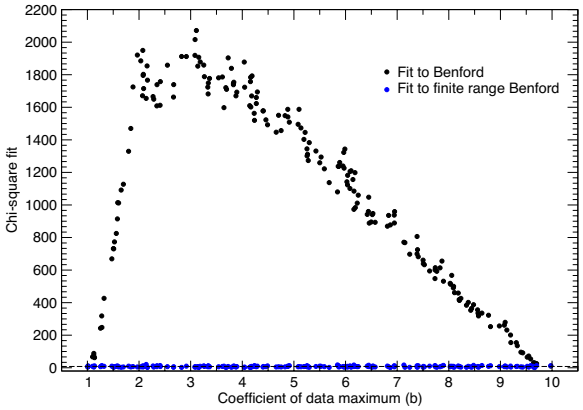

Fig. 2. Chi-square fit of first digit distributions to Benford predictions (black dots) given by (1) and the finite range theory (blue dots) given by (7) for 1000 sets of random deviates. The x axis is the upper coefficient b, the lower coefficient $a = 1$ and each data set has 10^5 deviates spanning three orders of magnitude, $\beta - \alpha = 3$.

where

$$\lambda_c = (\beta-\alpha) + \log_{10}\left(\frac{b}{a}\right) \tag{8}$$

and

$$\lambda_a = \begin{cases} \log_{10}\left(1+\frac{1}{D}\right) & : \quad D > a_1 \\ \log_{10}\left(\frac{1+D}{a}\right) & : \quad D = a_1 \\ 0 & : \quad D < a_1 \end{cases} \tag{9}$$

$$\lambda_b = \begin{cases} 0 & : \quad D > b_1 \\ \log_{10}\left(\frac{b}{D}\right) & : \quad D = b_1 \\ \log_{10}\left(1+\frac{1}{D}\right) & : \quad D < b_1. \end{cases} \tag{10}$$

Here a_1 is the first digit of a and b_1 is the first digit of b. At first sight this looks more complicated than the original Benford's law (1) but in fact it is no more difficult to evaluate. As we would expect (7)–(10) reduces to (1) when $a = 1$ or 10, and $b = 1$ or 10, and also as the dynamic range tends to infinity ($\beta - \alpha \to \infty$). By following similar arguments to that above (7) may be extended to the most general case of multiple digits (rather than one) and an arbitrary base, B. The result is almost identical to (7) except that all base 10 logarithms are replaced with logarithms in base B, and a_1 and b_1 are replaced with a_n and b_n indicating the first n digits of the data.

Figure 2 shows results of a numerical experiment comparing the fit of 1000 sets of random deviates to both Benford predictions given by (1) and the finite range version (7). A statistical chi-square measure is used to test goodness of fit of each first digit distribution to Benford law

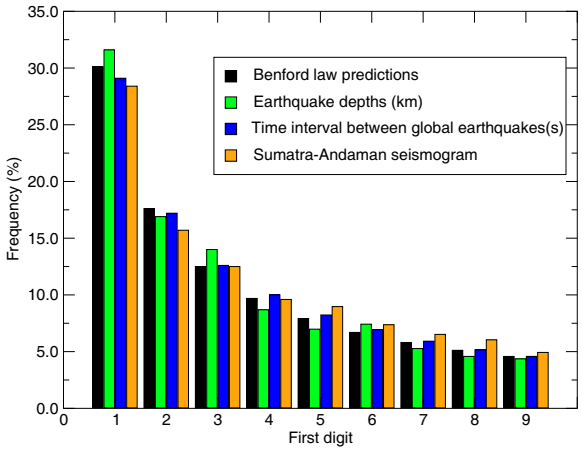

Fig. 3. Histograms of first digit distributions of three properties of earthquakes. These include 248,915 depths of globally distributed earthquakes in the period 1989–2009 (source National Earthquake Information Center, United States Geological Survey), separation times in seconds of 2,258,653 global earthquakes, and 24,000 ground displacements measured on a seismometer in Peru over the first 20 minutes of the Sumatra-Andaman earthquake of 2004.

predictions

$$\chi^2 = \sum_{D=1}^{9} \frac{(n_D - nP_D)^2}{nP_D}, \qquad (11)$$

where n_D is the observed frequency of digit D and there are n data in total. As the data range changes the chi-square goodness of fit is extremely poor for the standard Benford law and only obtains the expected value of eight when the deviate range is a whole number of decades, $b = 1$ and 10. In contrast the chi-square for the finite range prediction correctly fluctuates about the expected value of eight for all data sets. Similar trends are observed as the dynamic range of the data $(\beta - \alpha)$ are altered as well as number of deviates. In all cases, the predictions from (7) remain accurate.

4. Empirical Evidence in the Natural Sciences

Belatedly geophysicists have come across Benford's law and performed their own survey to test its presence in the physical sciences [8]. It was found that a wide range of modern phenomena appears to follow the digit law. In particular over 750,000 real numbers drawn from the fields of Physics, Astronomy, Geophysics, Chemistry, Engineering and Mathematics. These include the rotation frequencies of pulsars; green-house gas emissions, atmospheric temperature anomalies, masses of exoplanets, photon fluxes detected by the Fermi space telescope, as well as numbers of infectious diseases reported to the World Health Organisation. Random subsets of these data were also shown to fit Benford's law better than the original, which is consistent with predictions of the random samples theory of Hill [12].

5. Exploiting Benford's Law

To date the most practical use anyone has found for Benford's law is in a forensic mode, e.g. to detect fraud or rounding errors in real world numbers. This is possible by examining departures in the frequencies of individual digits from those predicted by Benford. This only makes sense once it is established (often empirically) that the data follow the law under normal circumstances. This has been exploited successfully to detect fraud in a range of situations involving financial data, such as business accounts, tax returns and stock market reports [23–25].

The successes in forensic accountancy have inspired physical scientists to see whether similar ideas might be applied to detect signals in contrast to background noise, e.g. in time series data. Recently geophysicists showed that Benford's law could be used to detect the onset of an earthquake from the frequencies of first digits of ground displacement counts recorded by a seismometer [8]. A summary of some collected results for digit distributions of earthquake properties is shown in Fig. 3 and includes their depth, time separation as well as displacements induced at the surface. Seismologists routinely identify and locate earthquakes using information from seismograms recorded at multiple locations across the globe. The ability to detect an earthquake from just the first digit histograms of seismic waveforms came as a surprise because most of the complex information contained in seismic waveforms would appear to have been removed when reducing the signal down to its first digit frequencies. Nevertheless the onset of an earthquake can clearly be manifested in the digit distribution alone. Inspired by this example, quantum physicists also applied Benford's law to the detection of quantum phase transitions with apparent success [26].

These examples show that analysis of digit frequencies has the potential to detect change in physical phenomena, which may lead to new

applications in the future. Seismologists have continued to examine the appearance of the digit law in their own field. A new data set that seems to follow Benford is the timing between earthquakes. Figure 3 shows results for over two million globally recorded earthquakes of all magnitudes and locations. The same result is achieved when earthquakes are divided into each magnitude range, except for the very large earthquakes (greater than magnitude eight) for which there are relatively few recorded in modern times. This result establishes a background trend for the digit frequencies of earthquake origin time separations. It raises the intriguing possibility that Benford's law might have applications in the detection of changes in the characteristics of earthquakes between regions, or over time. This may be an area for future exploitation.

6. Concluding Remarks

There is mounting evidence that Benford's law of first and later digit distributions may be a common feature across the physical sciences. While mathematicians will continue to seek theoretical justification for its existence, the challenge for physical scientists is to find new ways to exploit it. We argue that in situations where Benford's law is observed to hold empirically, localised departures from it are tell-tale features of some other process at play, perhaps worthy of more detailed investigation. This is in essence the same argument used successfully in forensic accountancy. As awareness of this novel phenomenon grows among physical scientists it seems likely that further applications will appear.

Over the years a large number of publications have appeared on aspects of Benford's law across multiple disciplines. This can make it difficult for newcomers to the subject to appreciate the range of results already known. Fortunately many publications are now accessible in a single online bibliography `http://www.benfordonline.net/` which is a valuable resource for all.

Acknowledgements

We thank Andrew Jackson and Ted Hill for useful discussions, and Rachel Fewster for pointing out her article [22].

References

[1] I. F. Goldstein and M. Goldstein, *The Experience of Science* (Plenum Press, New York, 1984).
[2] R. Snieder and K. Larner, *The Art of Being a Scientist: A Guide for Graduate Students and their Mentors* (Cambridge Univ. Press, Cambridge, 2009).
[3] I. P'. Lehmann, *Publications du Bureau Central Seismologique International* **A14** (1936) S.87–115.
[4] D. Clery, Bracing for a maelstrom of data, CERN puts its faith in the grid, *Science* **321** (2007) 1289–1291, doi: 10.1126/science.321.5894.1289.
[5] B. P. Abbott *et al.*, An upper limit on the stochastic gravitational-wave background of cosmological origin, *Nature* **460** (2009) 990–994.
[6] W. L. Jorgensen, The many roles of computation in drug discovery, *Science* **303** (2004) 1813–1818, doi: 10.1126/science.1096361.
[7] R. Lister *et al.*, Human dna methylomes at base resolution show widespread epigenomic differences, *Nature* **462** (2009) 315–332, doi:10.1038/nature08514.
[8] M. Sambridge, H. Tkalčić and A. Jackson, Benford's law in the natural sciences, *Geophys. Res. Lett.* **37** (2010) L22301, doi:10.1029/2010GL044830.
[9] S. Newcomb, Note on the frequency of use of different digits in natural numbers, *Amer. J. Math.* **4** (1881) 39–40.
[10] F. Benford, The law of anomalous numbers, *Proc. Am. Philos. Soc.* **78** (1938) 551–572.
[11] D. E. Knuth, *The Art of Computer Programming, Four Volumes* (Addison-Wesley, 1968).
[12] T. P. Hill, The first-digit phenomenon, *Amer. Sci.* **86** (1998) 358–363, doi: 10.1511/1998.4.358.
[13] E. Ley, On the peculiar distribution of the U.S. stock indexes' digits, *Amer. Stat.* **50** (1996) 311–314.
[14] B. F. Roukema, Benford's law anomalies in the 2009 Iranian presidential election, (2009), ArXiv e-prints 0906.2789.
[15] T. P. Hill, The significant-digit phenomenon, *Am. Math. Monthly* **102** (1995) 322–327.
[16] T. P. Hill, Base-invariance implies Benford's law, *Proc. Am. Math. Soc.* **123** (1995) 887–895.
[17] T. P. Hill, A statistical derivation of the significant-digit law, *Stat. Sci.* **10** (1995) 354–363.
[18] S. W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing* (California Technical Publishing, San Diego, CA, USA, 1997).
[19] A. Berger, Benford's law in power-like dynamical systems, *Stoch. Dyn.* **5** (2005) 587–607.
[20] W. Feller, *An Introduction to Probability Theory and Its Applications, Vol. II,* 2nd edition (John Wiley & Sons Inc., New York, 1971).
[21] A. Berger and T. Hill, Fundamental flaws in Feller's classical derivation of Benford's law, Univ. of Alberta preprint available from http://www.befordonline.net (2010).
[22] R. Fewster, A simple explanation of Benford's Law, *Amer. Stat.* **63** (2009) 26–32.
[23] M. J. Nigrini, The detection of income evasion through an analysis of digital distributions, Ph.D. thesis, Dept. of Accounting, University of Cincinnati (1992).
[24] M. J. Nigrini, A taxpayer compliance application of Benford's law, *J. Am. Tax Assoc.* **18** (1996) 72–91.
[25] M. J. Nigrini and S. J. Miller, Benford's law applied to hydrology data — results and relevance to other geophysical data, *Math. Geology* **39** (2007) 469–490.
[26] A. Sen De and U. Sen, Benford's law: detection of quantum phase transitions similarly as earthquakes (2011), arXiv:1103.5398v1 [quant-ph].

Malcolm Sambridge

Australian National University, Australia

Malcolm Sambridge obtained a PhD in Geophysics from the Australian National University in 1988. He spent periods as a post-doctoral research at the Carnegie Institution of Washington D C, USA, and University of Cambridge, UK. He currently leads the Seismology and Mathematical Geophysics group at the Australian National University. His research interests include data inference methods, inverse theory, seismology and earth structure, mathematical methods, Monte Carlo methods and optimisation.

Hrvoje Tkalčić

Australian National University, Australia

Hrvoje Tkalčić is a Fellow in Seismology at the Research School of Earth Sciences, the Australian National University. He holds a Diploma of Engineering Degree in Physics from the University of Zagreb, Croatia and PhD in Geophysics from the University of California at Berkeley. He studies structure and dynamics of Earth's interior using seismic waves generated by earthquakes, particularly its most inaccessible parts such as the inner and outer core. He is interested in increasing spatial coverage of the deep through the installations of seismometers in remote land areas and ocean bottoms.

Pierre Arroucau

North Carolina Central University, USA

Pierre Arroucau obtained his PhD in Seismology from the University of Nantes (France) in June 2006. After his PhD completion, he spent a year as a teaching and research assistant at the University of Nice-Sophia-Antipolis (France) before being awarded a postdoctoral fellowship at the Australian National University (Canberra, Australia) for two years. He is currently employed as a postdoctoral researcher at North Carolina Central University (Durham, NC, USA). His research interests include the acquisition, processing and interpretation of geophysical data.