European Journal of Physics

PAPER • OPEN ACCESS

Using residual heat maps to visualise Benford's multi-digit law

To cite this article: Benjamin Hull *et al* 2022 *Eur. J. Phys.* **43** 015803

OPEN ACCESS
IOP Publishing
European Journal of Physics
Eur. J. Phys. **43** (2022) 015803 (16pp)
https://doi.org/10.1088/1361-6404/ac3671

Using residual heat maps to visualise Benford's multi-digit law

Benjamin Hull[1], **Alexander Long**[1] **and Ifan G Hughes**[*]

Physics Department, Durham University, South Rd, Durham, DH1 3LE, United Kingdom

E-mail: i.g.hughes@durham.ac.uk

Received 21 July 2021, revised 13 October 2021
Accepted for publication 4 November 2021
Published 30 November 2021

Abstract

It has been known for more than a century that, counter to one's intuition, the frequency of occurrence of the first significant digit in a very large number of numerical data sets is nonuniformly distributed. This result is encapsulated in Benford's law, which states that the first (and higher) digits follow a logarithmic distribution. An interesting consequence of the counter intuitive nature of Benford's law is that manipulation of data sets can lead to a change in compliance with the expected distribution—an insight that is exploited in forensic accountancy and financial fraud. In this investigation we have applied a Benford analysis to the distribution of price paid data for house prices in England and Wales pre and post-2014. A residual heat map analysis offers a visually attractive method for identifying interesting features, and two distinct patterns of human intervention are identified: (i) selling property at values just beneath a tax threshold, and (ii) psychological pricing, with a particular bias for the final digit to be 0 or 5. There was a change in legislation in 2014 to soften tax thresholds, and the influence of this change on house price paid data was clearly evident.

Keywords: Benford's law, data analysis, data visualisation

(Some figures may appear in colour only in the online journal)

[*]Author to whom any correspondence should be addressed.
[1]These authors contributed equally to this work.

0143-0807/21/015803+16$33.00 Printed in the UK

1. Introduction

1.1. Benford's first digit law

Benford's first-digit law—also known as the Newcomb–Benford law, or the law of anomalous numbers—describes the counter intuitive phenomenon that the probability of the digits $1, 2, \ldots, 9$ to occur in the first index in a number drawn from many real-world data sets is not uniform [1]. Newcomb gave the first mathematical statement of the first-digit law after the observation that the first pages of logarithmic tables wear out faster than the last ones [2]. Over half a century later, Benford analysed 20 datasets, including the values of physical constants and population data, for conformity with Newcomb's findings. In addition to first-digit analysis, Benford's law can be extended to the second index, or any combination of indexes, in a number, such as the first-second digit Benford law [1].

Extensive numerical data sets exist that display non-uniform distribution of the first digit. For instance, the Fibonacci numbers, which appear in the growth patterns of sunflowers, pinecones and other plants and flowers [3]. The Fibonacci numbers are more likely to start with one than any other digit, with nine the least likely digit to appear. Indeed, the Fibonacci numbers are said to follow Benford's law [4] which quantifies this property. Numerous data sets that arise in different scientific disciplines have been found—to differing quantitative extents—to conform with Benford's law. The physics phenomena include alpha decay [5]; astrophysics [6]; atomic line strengths [7]; biophysics [8]; extra-solar planets [9]; geophysics [10]; particle physics [11]; quantum critical phenomena [12]; spectroscopy [13] and end-of-chapter exercises in physics textbooks [14]. An extensive database of articles, books and other resources related to Benford's law can be found at [15].

1.2. Benford's law and fraud detection

The counter intuitive nature of Benford's law has an interesting consequence: human intervention, or manipulation, of data sets can lead to a change in compliance with the first (or multi) digit law. In other words, many real-world data sets conform with Benford's law, except when the data are adjusted. This insight is used in forensic accountancy [16], and as a filter for financial and election fraud [17]. The central idea is that it should be easy to spot financial fraud as human intervention is unlikely to be conducted in a Benford-compliant way. Obviously, deviations from the expected Benford distribution do not necessarily have a nefarious origin, however they provide a useful flag for further investigation [16]. Consequently, there has been much interest in using Benford analysis for detecting anomalies in financial datasets [18–20].

In 1988 Carslaw [21] analysed the frequency of second digits occurring in reported earnings and losses for New Zealand based firms, concluding that there is a tendency to overstate earnings and understate losses. This study gave evidence of goal-oriented behaviour in financial statement reporting. These results were reproduced for firms operating in the United States and the United Kingdom by Thomas [22] and Van Caneghem [23], respectively. Tilden and Janes [24] showed that corporate accounting data conforms with Benford's law under stable economic conditions. Mehta [25] applies Benford's law, along with other forensic accounting techniques, to the financially fraudulent statements of Toshiba during the 2008–2015 period, concluding that Benford's law was a useful technique for detecting irregularities in reported accounting data in this instance. Previous studies of fraud using Benford's Law rarely, if ever, use multi-digit Benford analysis, despite it being shown that when financial manipulation occurs there is a tendency to round up profits [21] in order to meet company goals, something which multi-digit Benford analysis would excel at pointing out.

1.3. Benford's law and evidence of human intervention

One of the greatest difficulties with looking for evidence of human intervention in financial data sets is the lack of availability of a good control group. In this work we analyse price paid data as an example where a change of behaviour is expected. In England and Wales stamp duty tax is paid when purchasing a house, with different rates applied within certain price bands. There was a significant change in the way the tax was calculated in December 2014 (the higher rate being valid only for the excess amount above a threshold, versus applicable to the whole amount), therefore the data sets of prices paid for house sales pre and post 2014 is expected to be very different.

The simplest form of Benford's Law is a conformity test for first digit. A correction to the law to take into account the finite range of the data has been used. The strength, and weakness, of using a Benford filter is its simplicity; it is very easy to use for a large dataset. However, as noted above, one has to be very careful in interpretation of the (lack of) conformity of a data set with Benford's law. In this work we use multidigit tests, and introduce a finite range formula. We use heat maps of normalised residuals (difference between data set and expectation); this retains the ease of use of Benford's law, while also providing an easily visualised inspection method for evidence of human intervention.

In this work, we apply residual heat map analysis to house price stamp duty in England and Wales pre and post 2014, showing clear evidence of changes in behaviour of human intervention. From a university-level teaching perspective we emphasise that the techniques discussed and utilised in this paper are not restricted to financial datasets; indeed, they can be applied to a broad range of data-analysis problems. It can be enlightening for physics students to see techniques they master during their studies (chi-squared analysis; hypothesis compliance testing; quantitative analysis of differences between data and models) applied to a diverse range of real-world phenomena.

The rest of the paper is organised as follows: in section 2 we illustrate the use of the finite-range first digit law, discuss the multi-digit version, and introduce a finite-range correction to the later. In section 3 we present analysis of price paid data, showing in particular the power of the normalised residuals heat map to visualise trends in the data set. Finally, in section 4 we present our conclusions.

2. Benford distributions

In this section we shall look at the mathematical formulae for Benford's law, for both first and multi-digit form. The term index is used to refer to a position in a number. At each index there will be an integer $0, 1, 2, \ldots, 9$. Indexes begin at the first significant digit in a number. For example, for the integer 6301, the first index D_1 references the first position with a digit of 6, the second index D_2 references the second position, with a digit of 3 and similarly for the third and fourth indexes. In this case we would stay that $D_1 = 6, D_2 = 3, \ldots$ or $(D_1, D_2, D_3, D_4) = (6, 3, 0, 1)$. By assumption D_1 is non-zero.

2.1. First-digit distribution

Benford-compliant sets conform with a logarithmic distribution, with each successive digit appearing with a lower frequency in the first index. Newcomb derived the following formula describing the probability of the integer $d_1 \in \{1, 2, \ldots, 9\}$ to occur in the first index D_1,

$$P[D_1 = d_1] = \log\left(1 + \frac{1}{d_1}\right). \tag{1}$$

Note that all logarithms in this paper are to base 10, unless otherwise stated. Therefore according to Benford's law, the first digit of numbers appearing in Benford-compliant sets is more likely 1 (about 30%) than 2 (18%), and so forth to 9 (5%). There has been much discussion in the literature as to why such a large number of data sets should exhibit this logarithmic behaviour[2]; however, our interest in this work lies not in the fundamental reason behind the compliance (or not) of a particular dataset—there do not exist *a priori* criteria to know when a group of data should or should not fulfil the law [26]—but rather what can be learned by studying changes in compliance.

2.2. Finite-range version of the first-digit formula

Sambridge, Tkalčić, and Arroucau [27] formulated the first digit finite range version of Benford's law. Consider datasets in the range $U = [a \times 10^{\alpha}, b \times 10^{\beta}]$, $a, b \in [1, 10)$, $\alpha, \beta \in \mathbb{N}$ with $\beta > \alpha$. This form of Benford's law is useful when the upper and lower bounds of a dataset are known.

The probability of a number in the range U having a first digit $d_1 \in \{1, 2, \ldots, 9\}$ is given by the modified probability distribution function:

$$P[D_1 = d_1] = \frac{1}{\lambda_c}\left[(\beta - \alpha - 1)\log\left(1 + \frac{1}{d_1}\right) + \lambda_a + \lambda_b\right], \tag{2}$$

where,

$$\lambda_c = (\beta - \alpha) + \log\left(\frac{b}{a}\right)$$

$$\lambda_a = \begin{cases} \log\left(1 + \dfrac{1}{d_1}\right) & : d_1 > a_1 \\ \log\left(\dfrac{1 + d_1}{a}\right) & : d_1 = a_1 \\ 0 & : d_1 < a_1 \end{cases}$$

$$\lambda_b = \begin{cases} 0 & : d_1 > b_1 \\ \log\left(\dfrac{b}{d_1}\right) & : d_1 = b_1 \\ \log\left(1 + \dfrac{1}{d_1}\right) & : d_1 < b_1. \end{cases}$$

Here a_1 and b_1 are the first digits of a and b, respectively. The terms λ_a and λ_b in equation (2) are boundary terms accounting for data in the ranges $[a \times 10^{\alpha}, 10^{\alpha+1}]$ and $[10^{\beta}, b \times 10^{\beta}]$, respectively. The first term in the bracket is Benford's law applied to the range $[10^{\alpha+1}, 10^{\beta}]$. λ_c is a normalisation factor covering the entire finite range which ensures that all probabilities sum to one. A larger difference between α and β—the dynamic range of the data set—results in a better fit when applied to a Benford compliant set. Indeed, we recover Benford's law when $a, b = 1, 10$ or in the limit $\beta - \alpha \to \infty$.

[2] We would like to thank S J Clark for his insight that 'life is logarithmic'.

2.3. Quantification of compliance—normalised residuals

It is easy to glance at a histogram and tell whether the shape matches a Benford curve; indeed, such a simple approach is often adopted in popular accounts of Benford's law and some of the published literature. However, statistical tests are required to quantify the fit, and perhaps provide insight not apparent to the naked eye[3].

There are many possible tests and metrics for conformity of a data set with Benford's law that have been discussed in the literature. Each of these has its own strengths and weaknesses, and unfortunately there is no single metric that has been universally adopted. Our motivation in this work was to find a simple metric that can both be calculated easily and form the basis for a visual representation of the agreement (or not) of a data set with Benford's law; as we shall demonstrate below, this is particularly useful for the multi-digit law. Therefore, the test statistics used in this work are the widely used χ^2 and reduced χ^2 [28].

The χ^2 statistic for discrete data takes the form [28]

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}, \tag{3}$$

where E_i is the occurrence expected by the probability distribution, and O_i is the observed occurrence. In deciding how likely a value for χ^2 is to have occurred, the reduced χ^2 statistic, χ^2_ν, is useful:

$$\chi^2_\nu = \frac{\chi^2}{\nu}, \tag{4}$$

for data with ν degrees of freedom. A value of $\chi^2_\nu \approx 1$ implies a good match between the parent and sample distribution [28].

Note that the expected value E_i for having a first digit $d_1 \in \{1, 2, \ldots, 9\}$ is the product of the number of data points, N, with the probability from Benford's law, equation (2).

The χ^2 statistic can be thought of as the sum of the squares of the normalised residuals. The residual is defined as $O_i - E_i$, i.e. the difference between what we observe and what we expected; and a normalised residual is realised by dividing this value by its error bar, which is E_i for Poisson count statistics with discrete data [28]. Therefore for a data set that is expected to be compliant with Benford's law we would expect approximately two thirds of the normalised residuals to be less than 1 in magnitude.

We illustrate some of the points made above in figure 1 where we consider one of the original data sets form Benford's original paper, and the set of the first 1000 Fibonacci numbers (which are known to follow Benford's Law [29]).

2.4. Multi-digit distributions

The result of equation (1) can be generalised: the non-trivial positive integer $n = \sum_{i=0}^{m} 10^i a_i$ appears at the beginning of a number $N = \sum_{i=0}^{M} 10^i b_i$, with probability,

$$P[n] = \log\left(1 + \frac{1}{n}\right), \tag{5}$$

[3] Note that Benford's original paper did not use goodness-of-fit parameters to quantify conformity. Instead, Benford speculated on the 'probable error' on the data collected.

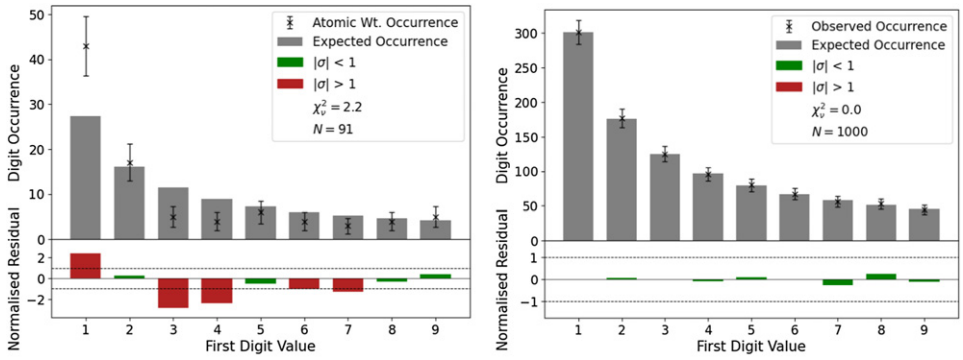


Figure 1. Histograms plotting (a) 'atomic wt.' data from Benford's paper [1] with 91 data points and (b) first digit data from the first 1000 Fibonacci numbers. The expected value of the occurrence, E_i, are in accordance with Benford's law, plotted as a cross, with error bars equal to the Poisson noise, $\sqrt{E_i}$. The actual observed occurrences, O_i, are shown for each digit as a bar. The digits for which the data are within a distance of one standard error of the value predicted by the Benford model are highlighted by their corresponding bars being green. Under the histograms the normalised residuals, $(O_i - E_i)/\sqrt{E_i}$, are plotted. Both data sets evidently display the first-digit phenomenon, with a far higher proportion of occurrences of 1 as the first digit than for any other number. However, for (a) the value of $\chi^2_\nu = 2.25$ indicating a poor conformity of atomic weights with Benford's law, whereas $\chi^2_\nu = 0.029$ for (b) as the Fibonacci numbers have a log-uniform distribution.

where $a_i, b_i \in \{0, 1, \ldots, 9\}$ and $a_m, b_M \neq 0$ with the decimal expansion of n having at most the same number of terms as N, that is, $m \leqslant M$.

A useful case of equation (5) is the first-second digit law which states that the probability of $d_1 \in \{1, 2, \ldots, 9\}$ occurring at the first index D_1 and $d_2 \in \{0, 1, \ldots, 9\}$ occurring at the second index D_2 is,

$$P[(D_1, D_2) = (d_1, d_2)] = \log\left(1 + \frac{1}{10d_1 + d_2}\right). \tag{6}$$

Note that $10d_1 + d_2$ is the decimal expansion of d_1 in the first index and d_2 in the second. Summing over all possible values of d_1 gives the probability of d_2 appearing in the second index D_2,

$$P[D_2 = d_2] = \sum_{d_1=1}^{9} \log\left(1 + \frac{1}{10d_1 + d_2}\right). \tag{7}$$

2.5. Finite-range version of the multi-digit formula

Here we present a finite-range version of Benford's multi-digit formula; the derivation of these formulae can be found in appendix 1. Consider a Benford set B in the range $[a \times 10^\alpha, b \times 10^\beta]$, with $\alpha > \beta$ integers and $a, b \in \mathbb{R}^{\geqslant 1}$. Then the probability of the n digit long integer $D = \sum_{i=1}^{n} d_i \times 10^{n-i} = d_1 \times 10^{n-1} + d_2 \times 10^{n-2} + \cdots + d_n$ to appear at the beginning of an

element $x \in B$, $P_D(n)$, is given by,

$$P_D(n) = \frac{1}{\lambda_c}\left[(\beta - \alpha - 1)\log\left(1 + \frac{1}{D}\right) + \lambda_a + \lambda_b\right], \tag{8}$$

where

$$\lambda_c = (\beta - \alpha) + \log\left(\frac{b}{a}\right) \tag{9}$$

and

$$\lambda_a = \begin{cases} \log\left(1 + \dfrac{1}{D}\right) & D > a_n \\ \log\left(\dfrac{1+D}{a \times 10^{n-1}}\right) & D = a_n \\ 0 & D < a_n \end{cases} \tag{10}$$

$$\lambda_b = \begin{cases} 0 & D > b_n \\ \log\left(\dfrac{b \times 10^{n-1}}{D}\right) & D = b_n \\ \log\left(1 + \dfrac{1}{D}\right) & D < b_n. \end{cases} \tag{11}$$

Here a_n and b_n are the first n digits of a and b respectively. If a or b is less than n digits in length then assume all subsequent digits are zero. Note that $d_i \in \{0, 1, \ldots, 9\}$ and $d_1 \in \{1, 2, \ldots, 9\}$.

Figure 2 illustrates the concepts introduced in this section. We use the (a) conventional, equation (5), and (b) finite-range, equation (8), first-second digit Benford's law to analyse a geometric series (which are known to be Benford compliant [29]). We plot the normalised residuals as a heat map for all 90 possible first-second digit combinations. The geometric series analysed had an upper bound of $\approx 4.3 \times 10^9$. A sudden shift from positive residuals to negative residuals appears immediately after the $(d_1, d_2) = (4, 3)$ point; the finite-range law, on the other hand, does not exhibit a sharp feature in its residuals, indicative of a more uniform conformity.

3. House price data

We now move on to apply the ideas and formulae developed above to a specific worked example—house price paid data from England and Wales [30]. We shall consider the data in two distinct sets, pre and post 2014, for reasons outlined below.

3.1. Stamp duty tax

A notable factor that may influence house price values is stamp duty tax—a levy that applies to property transactions in England and Wales, and is payable to HM Revenues and Customs. The tax rates depends on the price of the property, i.e. the tax bracket of the property. The tax rate and brackets pre and post December 2014 are outlined in table 1.

3.2. Psychological pricing

It has been noted previously [32–37] that human intervention is setting commodity prices leads to certain preferred values—this is at odds with Benford's law, and we therefore expect our

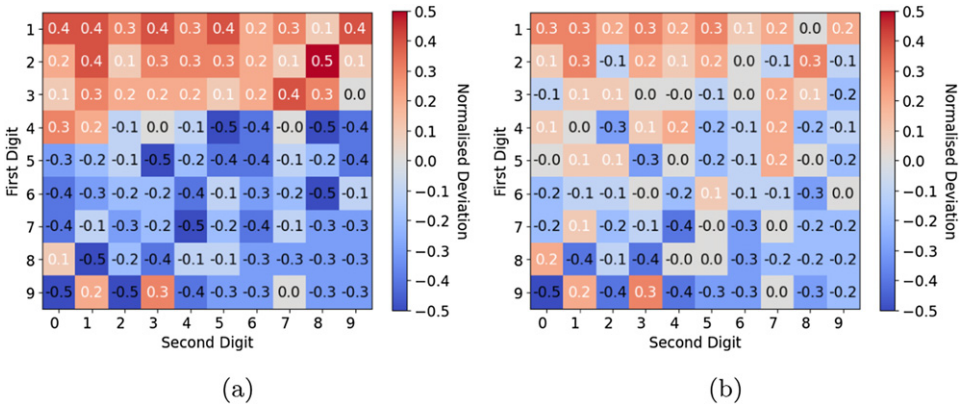


Figure 2. First-second digit normalised residuals of a geometric series with 2,000 data points compared to the (a) regular first-second digit Benford's law and (b) the finite-range first-second digit Benford's law. The geometric series' upper bound is $\approx 4.3 \times 10^9$, and a sudden shift from positive residuals to negative residuals appears just after the $(d_1, d_2) = (4, 3)$ point; the finite-range law on the other hand does not feature a sharp turning point of its residuals.

Table 1. Threshold values and stamp duty tax rates for residential properties valued under one million pounds sold in England and Wales [31]. Pre December 2014 the rate was payable on the total value of the property (the firm rate) whereas post December 2014 the rate was payable only on the portion of the value of the property in each tax bracket (the relaxed rate). Note that some of the tax brackets change between the two rates.

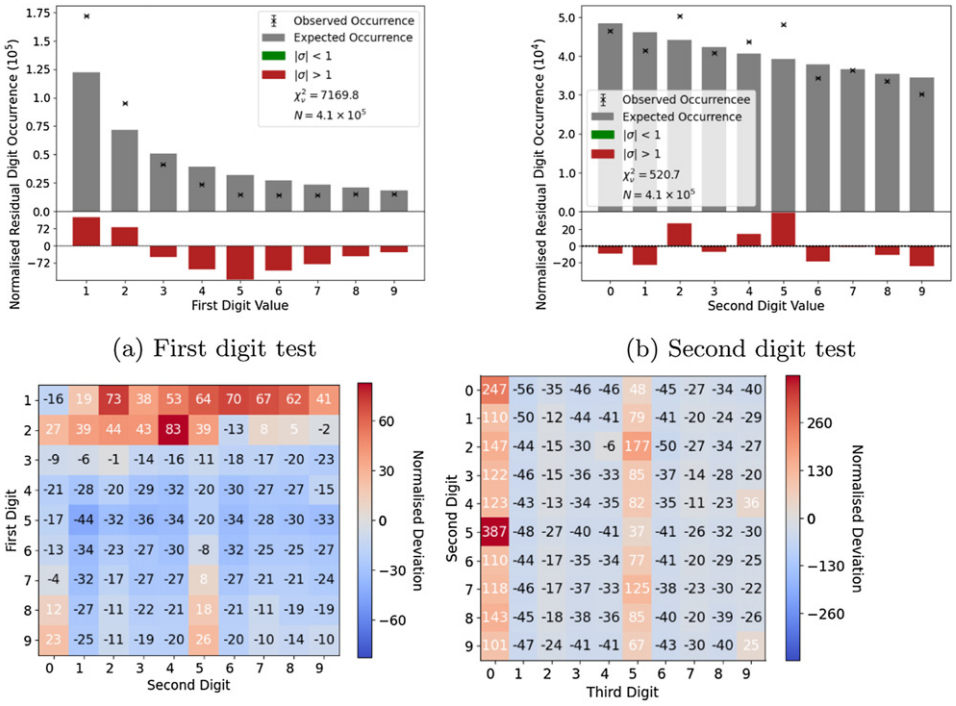
Tax rate after Dec. 2014 (relaxed)		Tax rate pre Dec. 2014 (firm)	
Tax bracket	Tax rate (%)	Tax bracket	Tax rate (%)
£0–£125 000	0	£0–£125 000	0[a]
£125 001–£250 000	2	£125 001–£250 000	2[a]
£250 001–£925 000	5	£250 001–£500 000	3[a]
£925 001–£1.0m	10	£500 000–£1.0m	4[a]

[a]Payable on total property price once limit is breached.

Benford multi-digit filter to reveal this *psychological pricing* behaviour. There is a tendency to price houses at given reference points. Namely, we observe that property prices tend to be set with zero or five in the second and third digit; i.e. a property is more likely to be priced at £250k or £255k than £252 500. This is similar to the cognitive reference points in commodity pricing described by Carslaw [21] in that there is a psychological tendency for properties to be valued at these price points. We shall also demonstrate that there is a tendency to price houses just below psychological price points in agreement with the pricing behaviour described by Rosch [38]; i.e. there is a human perception that £99k is significantly cheaper than £100k.

3.3. Price paid data pre and post 2014

Figure 3 shows the results of looking for conformity between Benford's law and house price paid data for 2013 and 2014. Subfigures (a) and (b) show the first and second digit tests,



(a) First digit test

(b) Second digit test

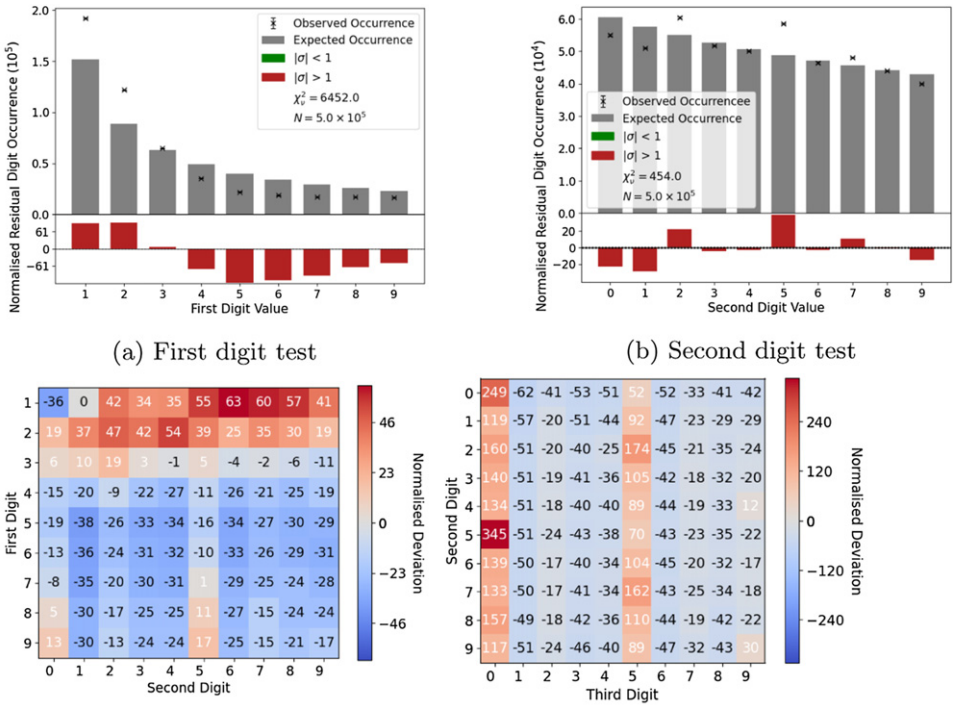
(c) First-second digit heat map showing normalised deviation from Benford's law

(d) Second-third digit heat map showing normalised deviation from Benford's law

Figure 3. Subfigures (a) and (b) show the expected Benford occurrence in the first and second indexes, respectively, and the observed occurrence for price paid data for houses in England and Wales in the years 2013 and 2014, with 4.1×10^5 data points. Subfigures (c) and (d) are heat maps showing the difference between the observed and expected Benford occurrence for the first-second and second-third index pairs, respectively. The heat maps are normalised with respect to the standard error. Red colouration corresponds to a higher than expected occurrence and blue a lower than expected occurrence. In (a) we note a clear tendency for properties to be priced with one or two in the first index. Moreover, properties tend to be valued with five in the second digit, which can be seen in middle column corresponding to a second digit of five in (c) and the normalised residual at the digits five in (b). (d) shows that properties are priced most frequently with zero or five in the third index, as evidenced by the orange columns corresponding to third digits of zero and five. The most common second-third digit pairs are (0, 0) and (5, 0).

respectively. Each of these tests suggests nonconformity with a Benford distribution at the 0.01 confidence level according to the χ^2 test statistics. This conclusion comes as no surprize, as it is evident that all the first and second digits deviate from expectation by more than one normalised error. For the first digit test, the distribution is skewed towards one and away from nine, while the second digit test shows that five is the most likely digit to appear in the second index.

Similarly, figure 4 shows a Benford analysis for price paid data from 2015 and 2016. Again, subfigures (a) and (b) show the first and second digit tests, respectively. As before, we observe nonconformity at the 0.01 confidence level. All data points deviate from expectation by more than one normalised residual for both the first and second digit tests. The distribution of first



(a) First digit test

(b) Second digit test

(c) First-second digit heat map showing normalised deviation from Benford's law

(d) Second-third digit heat map showing normalised deviation from Benford's law

Figure 4. Subfigures (a) and (b) show the expected Benford occurrence in the first and second indexes, respectively, and the observed occurrence for price paid data for houses in England and Wales in 2015 and 2016, with 5.0×10^5 data points. Subfigures (c) and (d) are heat maps showing the difference between the observed and expected Benford occurrence for the first-second and second-third index pairs, respectively. The heat maps are normalised with respect to the standard error. Red colouration corresponds to a higher than expected occurrence, and blue a lower than expected occurrence. Similar to figure 3, in (a) there is clear evidence suggesting houses are priced with one or two in the first index, as indicated by a high concentration of red values in the first two normalised residuals. Once again, houses tend to be valued with five in the second digit, which can be seen in the middle column corresponding to a second digit of five in (c) and the normalised residual at the digit five in (b). As in 2013, (d) shows that houses are frequently priced with zero or five in the third index, as evidenced by the red columns corresponding to third digits of zero and five. Again the most common second-third digit pairs are (0, 0) and (5, 0), emphasising the financial and psychological factors affecting property valuations.

digits is skewed towards one, and five deviates the most from the expected occurrence of second digits.

From a purely traditional Benford analysis standpoint, it is difficult to distinguish the two distributions. Both are highly nonconforming and, upon casual inspection, appear to show the same trends. However, the heat maps in subfigures (c) and (d) in figures 3 and 4 reveal subtle variations between the two.

Subfigures (c) show the normalised deviation of the first and second digits from a Benford distribution. For the 2013 data, shown in figure 3(c), we expect properties to be valued just

below the tax threshold values shown in table 1. Indeed there is a spike for the first and second digits $(D_1, D_2) = (1, 2), (2, 4)$ which correspond to valuations just below the threshold values of £125 001 and £250 001. We observe a similar, yet less pronounced, result in figure 4(c). This is most likely due to the relaxed tax rates, as there is less motivation to fix prices below threshold values (see table 1).

Not all the first-second digit values $(1, 2)$ occur in the range £120k–£125k. Indeed for the 2013 data, 52.4% of all data points in the range £120k–£130k lie in the range £120k–£125k compared with 48.5% for the 2015 data. This suggests that there is more of a tendency to value properties under the threshold value for valuations made under firm tax rates compared to the relaxed rates.

We observe pricing at the reference points £120k and £125k in the range £120–£125. Indeed 17.7% and 18.4% of properties are priced at £120k in 2013 and 2015, respectively, in this price range. Moreover, 31.0% and 23.4% are priced at £125k in 2013 and 2015, respectively, showing a clear tendency for valuations to be made with zero or five in the third digit. This psychological bias is shown in figures 3(d) and 4 (d), which show the normalised deviation of second and third digit occurrence from Benford's law. Indeed, the second-third digit pairs (5, 0) and (0, 0) display the most deviation, showing the significance of this property-pricing factor.

There is also a tendency to price houses just below psychological price points in agreement with the pricing behaviour described by Rosch [38]. In particular, in figures 3(d) and 4 (d) we see spikes at the second-third digit pairs (4, 9) and (9, 9). Such price points correspond to prices set just below the psychological prices with five and zero in the second digit. This is done to give the impression of a significantly lower price than the property's actual price. That is a house priced at £249k appears to be significantly cheaper than if it were priced at £250k. This is a well-known technique used in sales and is used when marketing properties.

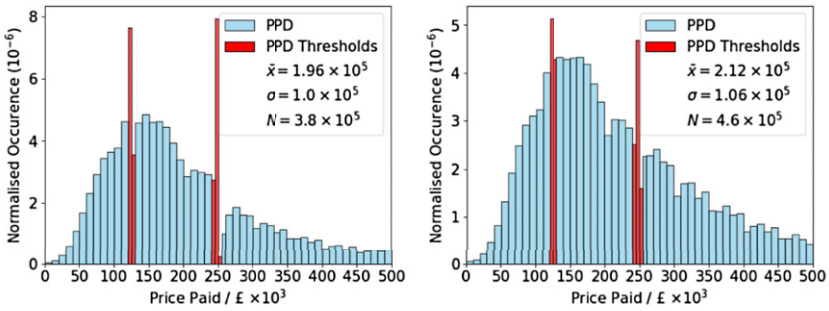
Figure 5 shows histograms of the distribution of price paid data in the years (a) 2013 and (b) 2015. The bins were chosen beginning at one with a width of ten thousand, except around tax thresholds, in which instance they have a width of five thousand. Such a choice allows us to examine properties priced at £125k and £250k.

In 2013 there was a clear pattern of pricing properties to avoid higher taxation, particularly at the £250,001 threshold, which incurs a 3% tax rate on the entire property. Figure 3(c) shows the same features, as there is a higher count of the first-second digits (1, 2) than expected, corresponding to the £125,001 threshold and an increased count of (2, 4) corresponding to the £250 001 threshold. Similar, yet less prominent, features hold in 2015 primarily due to relaxed tax rates. Once again figure 4(c) displays the same trends with spikes in the first-second digits (1, 2) and (2, 4); however, these features are less significant when compared to figure 3(c).

We note that with the exception of the tax thresholds there appears to be a smooth distribution to describe the form of the histogram. We emphasise again that the exact form of this distribution is neither particularly interesting nor relevant for this investigation; it is the pattern of the deviation from the smooth distribution that is analysed here.

Although there is a clear trend in valuing properties below tax threshold values, such valuations may also be influenced by psychological bias. That is, properties may be valued at £125k or £250k since there is an aforementioned bias towards pricing properties with a zero or five in the third index. Although there is a clear motivation to set prices at this value to maximise the profit made while selling the property and minimising tax paid, it is unclear whether this is done deliberately in every case or instead, whether this behaviour can be attributed to unconscious bias. It should also be noted that there is no mechanism for checking this within the scope of this investigation.

The results of first-digit analyses for both the 2013 and 2015 PPD can be explained using figure 5. For example, in 2015, 62% of data points are in the range £100k–£299k. Indeed,



(a) 2013 price paid data for houses in England and Wales showing threshold value drop-off

(b) 2015 for price paid data for houses in England and Wales showing threshold value drop-off

Figure 5. Histograms showing the distribution of price paid data for houses in England and Wales in the years 2013 (3.8×10^5 data points) and 2015 (4.6×10^5 data points). Bars shown in red represent prices near the tax threshold values £125 001 and £250 001. The legend shows the mean $\bar{x}$ and standard deviation σ of house prices as well as the number of house prices analysed N. There is clear pattern of houses been priced below these tax thresholds corresponding to spikes in the histograms just below these values and deficits immediately after. Indeed, this trend is most pronounced in subfigure (a) under the firm tax rate in 2013, suggesting valuations are made with tax minimisation taken into consideration. Similar features hold in 2015 under the relaxed tax rate; however, they are less pronounced.

inspecting figure 5 reveals that most property prices are in the range £100k–£299k. Thus we would expect deviations from Benford's law due to this concentration of data points close to the mean (£212k). A similar argument holds for the 2013 price paid data.

4. Conclusions

In this investigation we have applied a Benford analysis to the distribution of price paid data for house prices in England and Wales pre and post-2014. Neither distribution conforms with Benford's law, and they appear on face value to display roughly the same trends. Although a traditional Benford analysis reveals this nonconformity, there is little insight indicating why this should be the case. A residual heat map analysis for both first-second and second-third digits allowed further insight to be gained. Two examples of human intervention on price paid data were revealed; (i) selling property at values just cheaper than a tax threshold, to reduce the amount of tax paid; and (ii) psychological pricing, with a particular bias for the final digit to be 0 or 5. There was a change in legislation in 2014 to soften the presence of tax thresholds, and the influence of this change on house price paid data was clearly evident.

We note that income tax in the UK also has different rates within different brackets. However, it is significantly easier to obtain the data set for price paid data for property than it is for income tax, as individual tax returns are subject to privacy and confidentiality issues [39]. Nevertheless, a multi-digit heat map Benford analysis of income tax data could be an interesting area of future investigation.

The techniques utilised in this work are simple to apply and efficient to use. The residual heat map analysis in particular offers a visually attractive method for identifying interesting features in a large data set, such as human intervention. However, we finish with a note of caution: while

these techniques provide a fast method of analysing massive data sets for anomalous behaviour, as Goodman points out in the context of financial fraud detection [40], Benford's law is not a hypothesis test, and no individual should be accused of fraud based solely on evidence from Benford analysis.

Acknowledgments

We thank Steven Wrathmall for useful discussions and advice on the manuscript, and Martin Ward and Charlotte Wojcik for suggesting the topic of house price stamp duty as a rich source of data for a Benford analysis.

Appendix A. Finite range law proof

We will assume the PDF has a log-uniform distribution or equivalently $P(x) = x^{-1}$. This assures that the underlying set is Benford [27]. Write a_n as $a_n = \sum_{i=1}^{n} a_i \times 10^{n-i} = a_1 \times 10^{n-1} + a_2 \times 10^{n-2} + \cdots + a_n$ and similarly for b_n.

We will consider three subsets of B defined by $L := [a \times 10^{\alpha}, 10^{\alpha+1}]$, $M := [10^{\alpha+1}, 10^{\beta}]$ and $R := [10^{\beta}, b \times 10^{\beta}]$. Note that M can be empty if $\beta = \alpha + 1$ but one of L and R will be non-empty.

Firstly, we can determine the normalisation constant, λ_c, by integrating the probability density $P(x) = x^{-1}$ over the entire range of the set.

$$\begin{aligned}\lambda_c &= \int_{a\times 10^{\alpha}}^{b\times 10^{\beta}} \frac{1}{x}\,\mathrm{d}x = (\beta-\alpha)\ln(10) + \ln\left(\frac{a}{b}\right)\\ &= \frac{1}{\ln(10)}\left[(\beta-\alpha) + \log\left(\frac{a}{b}\right)\right].\end{aligned}$$

$\underline{M = [10^{\alpha+1}, 10^{\beta}]:}$ Define $M^{\alpha+i} := [10^{\alpha+i}, 10^{\alpha+i+1}]$ and the set $M_D^{\alpha+i}$ as the subset of $M^{\alpha+i}$ with D as the first n digits. Then,

$$M_D^{\alpha+i} = [D \times 10^{\alpha+i-n-1}, (D+1) \times 10^{\alpha+i-n-1}).$$

The integral of the density function over all $M_D = \bigcup_{i=1}^{\beta-\alpha-1} M_D^{\alpha+i}$ is,

$$\begin{aligned}\sum_{i=1}^{\beta-\alpha-1} \int_{M_D^{\alpha+i}} \frac{1}{x}\,\mathrm{d}x &= \sum_{i=1}^{\beta-\alpha-1} \int_{D\times 10^{\alpha+i-n-1}}^{(D+1)\times 10^{\alpha+i-n-1}} \frac{1}{x}\,\mathrm{d}x\\ &= \sum_{i=1}^{\beta-\alpha-1} \ln\left(1+\frac{1}{D}\right)\\ &= \frac{1}{\ln(10)}\left[(\beta-\alpha-1)\log\left(1+\frac{1}{D}\right)\right].\end{aligned}$$

$\underline{L := [a \times 10^{\alpha}, 10^{\alpha+1}]:}$ The set of elements in L with D as the first n digits is given by,

$$L_D = \begin{cases} \left[D \times 10^{\alpha-n-1}, (D+1) \times 10^{\alpha-n-1}\right] & : D > a_n,\\ \emptyset & : D < a_n,\\ [a \times 10^{\alpha}, (D+1) \times 10^{\alpha-n-1}] & : D = a_n.\end{cases}$$

We then integrate over the density function to obtain the coefficient λ_a:

$$\lambda_a := \int_{L_D} \frac{1}{x}\,\mathrm{d}x = \begin{cases} \displaystyle\int_{D\times 10^{\alpha-n-1}}^{(D+1)\times 10^{\alpha-n-1}} \frac{1}{x}\,\mathrm{d}x & : D > a_n, \\ 0 & : D < a_n, \\ \displaystyle\int_{a\times 10^{\alpha}}^{(D+1)\times 10^{\alpha-n-1}} \frac{1}{x}\,\mathrm{d}x & : D = a_n, \end{cases}$$
$$= \begin{cases} \dfrac{1}{\ln(10)}\,\log\left(1+\dfrac{1}{D}\right) & : D > a_n, \\ 0 & : D < a_n \\ \dfrac{1}{\ln(10)}\,\log\left(\dfrac{D+1}{a\times 10^{n-1}}\right) & : D = a_n. \end{cases}$$

$R := [10^{\beta}, b\times 10^{\beta}]$: The set of elements in R with D as the first n digits is given by,

$$R_D = \begin{cases} \emptyset & : D > b_n, \\ [D\times 10^{\beta-n-1}, (D+1)\times 10^{\beta-n-1}] & : D < b_n, \\ [D\times 10^{\beta-n-1}, b\times 10^{\alpha}] & : D = b_n. \end{cases}$$

We then integrate over the density function to obtain the coefficient λ_b:

$$\lambda_b := \int_{R_D} \frac{1}{x}\,\mathrm{d}x = \begin{cases} 0 & : D > b_n, \\ \displaystyle\int_{D\times 10^{\beta-n-1}}^{(D+1)\times 10^{\beta-n-1}} \frac{1}{x}\,\mathrm{d}x & : D < b_n, \\ \displaystyle\int_{D\times 10^{\beta-n-1}}^{b\times 10^{\beta}} \frac{1}{x}\,\mathrm{d}x & : D = b_n, \end{cases}$$
$$= \begin{cases} 0 & : D > b_n, \\ \dfrac{1}{\ln(10)}\,\log\left(1+\dfrac{1}{D}\right) & : D < b_n \\ \dfrac{1}{\ln(10)}\,\log\left(\dfrac{b\times 10^{n-1}}{D}\right) & : D = b_n. \end{cases}$$

Summing each of these integrals over L, M and R and dividing by the normalisation constant, λ_c, gives the probability of finding an element in the Benford set in the range $[a\times 10^{\alpha}, b\times 10^{\beta}]$ with the first n digits being D.

$$\begin{aligned} P_n(D) &= \frac{1}{\int_{a\times 10^{\alpha}}^{b\times 10^{\beta}} \frac{1}{x}\,\mathrm{d}x}\left[\int_{M_D}\frac{1}{x}\,\mathrm{d}x + \int_{L_D}\frac{1}{x}\,\mathrm{d}x + \int_{R_D}\frac{1}{x}\,\mathrm{d}x\right] \\ &= \frac{1}{\lambda_c}\left[\frac{(\beta-\alpha-1)}{\ln(10)}\,\log\left(1+\frac{1}{D}\right) + \lambda_a + \lambda_b\right]. \end{aligned} \tag{12}$$

Note that we can cancel the $\frac{1}{\ln(10)}$ terms from all terms in the bracket with the $\frac{1}{\ln(10)}$ in λ_c which gives the desired form.

ORCID iDs

Alexander Long https://orcid.org/0000-0001-5294-0334
Ifan G Hughes https://orcid.org/0000-0001-6322-6435

References

[1] Benford F 1938 The law of anomalous numbers *Proc. Am. Phil. Soc.* **78** 553 https://jstor.org/stable/984802
[2] Newcomb S 1881 Note on the frequency of use of the different digits in natural numbers *Am. J. Math.* **4** 39–40
[3] Omotehinwa T and Ramon S 2013 Fibonacci numbers and golden ratio in mathematics and science *Int. J. Comput. Inf. Technol.* **2** 630–8
[4] Kunoff S 1987 N! has the first digit property *Fibonacci Q.* **25** 365–7
[5] Buck B, Merchant A C and Perez S M 1993 An illustration of Benford's first digit law using alpha decay half lives *Eur. J. Phys.* **14** 59–63
[6] de Jong J, de Bruijne J and De Ridder J 2020 Benford's law in the gaia universe *Astron. Astrophys.* **642** A205
[7] Pain J-C 2008 Benford's law and complex atomic spectra *Phys. Rev.* E **77** 012102
[8] da Silva A, Floquet S, Santos D and Lima R 2020 On the validation of the Newcomb–Benford law and the Weibull distribution in neuromuscular transmission *Physica* A **553** 124606
[9] Shukla A, Pandey A K and Pathak A 2017 Benford's distribution in extrasolar world: do the exoplanets follow Benford's distribution? *J. Astrophys. Astron.* **38** 7
[10] de Macedo I A S and de Figueiredo J J S 2018 Using Benford's law on the seismic reflectivity analysis *Interpretation* **6** T689–97
[11] Shao L and Ma B-Q 2009 First digit distribution of hadron full width *Mod. Phys. Lett.* A **24** 3275–82
[12] Bera A, Mishra U, Singha Roy S, Biswas A, Sen(De) A and Sen U 2018 Benford analysis of quantum critical phenomena: first digit provides high finite-size scaling exponent while first two and further are not much better *Phys. Lett.* A **382** 1639–44
[13] Bhole G, Shukla A and Mahesh T S 2015 Benford analysis: a useful paradigm for spectroscopic analysis *Chem. Phys. Lett.* **639** 36–40
[14] Slepkov A D, Ironside K B and DiBattista D 2015 Benford's law: textbook exercises and multiple-choice testbanks *PLOS ONE* **10** e0117972
[15] Berger A, Hill T P and Rogers E 2009 Benford online bibliography https://benfordonline.net/ (accessed 01 July 2021)
[16] Parker M 2020 *Humble Pi: A Comedy of Maths Errors* (London: Penguin)
[17] Roukema B F 2014 A first-digit anomaly in the 2009 iranian presidential election *J. Appl. Stat.* **41** 164–99
[18] Nigrini M 2005 An assessment of the change in the incidence of earnings management around the Enron-Andersen episode *Rev. Account. Finance* **4** 92–110
[19] Nigrini M J 2012 *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection* vol 586 (New York: Wiley)
[20] Diekmann A and Jann B 2010 Benford's law and fraud detection: facts and legends *German Econ. Rev.* **11** 397–401
[21] Carslaw C 1988 Anomalies in income numbers: evidence of goal oriented behavior *Account. Rev.* **63** 321–7 https://jstor.org/stable/248109
[22] Thomas J K 1989 Unusual patterns in reported earnings *Account. Rev.* **64** 773–87 https://jstor.org/stable/247861?seq=1
[23] Van Caneghem T 2002 Earnings management induced by cognitive reference points *British Account. Rev.* **34** 167–78

[24] Tilden C and Janes T 2012 Empirical evidence of financial statement manipulation during economic recessions *J. Finance Account.* **12** 1–15 http://aabri.comwww.aabri.com/manuscripts/121125.pdf
[25] Mehta A and Bhavani G 2017 Application of forensic tools to detect fraud: the case of Toshiba *J. Forensic Investigative Account.* **9** 1188–97
[26] Torres J, Fernández S, Gamero A and Sola A 2007 How do numbers begin? (the first digit law) *Eur. J. Phys.* **28** L17–25
[27] Sambridge M, Tkalčić H and Arroucau P 2011 Benford's law of first digits: from mathematical curiosity to change detector *Asia Pacific Math. Newslett.* **1** 1–6
[28] Hughes I G and Hase T P A 2010 *Measurements and Their Uncertainties: A Practical Guide to Modern Error Analysis* (Oxford: Oxford University Press)
[29] Raimi R A 1976 The first digit problem *Am. Math. Monthly* **83** 521–38
[30] HM Land Registry 2020 Archived Price Paid Data: 1995 to 2017 https://data.gov.uk/dataset/314f77b3-e702-4545-8bcb-9ef8262ea0fd/archived-price-paid-data-1995-to-2017 (accessed 06 January 2021)
[31] The National Achieves 2015 Stamp duty land tax act 2015 [Online] https://legislation.gov.uk/ukpga/2015/1/contents (accessed 07 January 20)
[32] Kreul L M 1982 Magic numbers: psychological aspects of menu pricing *Cornell Hotel Restaur. Adm. Q.* **23** 70–5
[33] Basu K 1997 Why are so many goods priced to end in nine? and why this practice hurts the producers *Econ. Lett.* **54** 41–4
[34] Wedel M and Leeflang P S 1998 A model for the effects of psychological pricing in Gabor–Granger price studies *J. Econ. Psychol.* **19** 237–60
[35] Aggarwal R and Lucey B M 2007 Psychological barriers in gold prices? *Rev. Financ. Econ.* **16** 217–30
[36] Levy D, Snir A, Gotler A and Chen H A 2020 Not all price endings are created equal: price points and asymmetric price rigidity *J. Monetary Econ.* **110** 33–49
[37] Hillen J 2021 Psychological pricing in online food retail *British Food J.* **123** 3522–35
[38] Rosch E 1975 Cognitive reference points *Cogn. Psychol.* **7** 532–47
[39] Wojcik C 2020 Can benford's law be used to detect financial fraud? *Master's Thesis* Department of physics, Durham University
[40] Goodman W 2016 The promises and pitfalls of Benford's law *Significance* **13** 38–41