The current issue and full text archive of this journal is available on Emerald Insight at:
www.emeraldinsight.com/1359-0790.htm

Assessing the omission of records from a data set using Benford's law

Pedro Carreira
School of Technology and Management and CIGS Management for Sustainability Research Center, Leiria Polytechnic Institute, Leiria, Portugal, and

Carlos Gomes da Silva
School of Technology and Management, Leiria Polytechnic Institute, Leiria, Portugal and INESC Coimbra, Coimbra, Portugal

Abstract

Purpose – The purpose of this paper is to propose a methodology to estimate the number of records that were omitted from a data set, and to assess its effectiveness.

Design/methodology/approach – The procedure to estimate the number of records that were omitted from a data set is based on Benford's law. Empirical experiments are performed to illustrate the application of the procedure. In detail, two simulated Benford-conforming data sets are distorted and the procedure is then used to recover the original patterns of the data sets.

Findings – The effectiveness of the procedure seems to increase with the degree of conformity of the original data set with Benford's law.

Practical implications – This work can be useful in auditing and economic crime detection, namely in identifying tax evasion.

Originality/value – This work is the first to propose Benford's law as a tool to detect data evasion.

Keywords Financial crime, Auditing, Benford's law, Digital analysis, Fraud detection

Paper type Research paper

1. Introduction

The decisions made by economic agents are based on the disclosed and exchanged information between them. By omitting some information from a data set (as, for example, when some sales are unrecorded by firms), economic agents harm the functioning of the economy. In the fiscal domain, evasion practices may lead to decreased government revenue that could be used to raise global welfare. Also, at a financial level, biased information may lead economic agents to invest more frequently in undesirable projects or to neglect profitable ones.

The development of tools to audit and monitor the authenticity of high flows of numerical information, especially those supported by information technologies, is thus of relevance. Digital analysis, defined as the study of digits and patterns of numbers (Nigrini and Mittermaier, 1997), has proved to be useful with this respect, as its underlying techniques (which can be found for example in Coderre, 2009) can easily deal with huge amounts of data, while being simple and automated.

A particular digital analysis technique consists of the application of Benford's law (Newcomb, 1881; Benford, 1938), which can be used to detect misbehavior of samples of

Journal of Financial Crime
Vol. 23 No. 4, 2016
pp. 798-805
© Emerald Group Publishing Limited
1359-0790
DOI 10.1108/JFC-10-2015-0060

variables from many different fields, such as economics, physics or accounting, for example. References on the application of Benford's law in auditing include Carlslaw (1988), Nigrini (1994), Nigrini and Mittermaier (1997), Drake and Nigrini (2000), Nigrini (2000), Durtschi *et al.* (2004), Diekmann (2007), Watrin *et al.* (2008), Coderre (2009) and Gomes da Silva and Carreira (2013).

When it can be applied (see Hill, 1995, for the theoretical requirements, and Durtschi *et al.*, 2004, for examples of variables that must follow the law), Benford's law states that the probability (e_i) of a number having first digit i follows a logarithmic function given by $e_i = \log(1 + 1/i)$, $i = 1,\ldots,9$. This function gives also the probabilities of higher orders for the first digits of a number, such as, for example, the first two and first three digits, in which cases $i = 10,\ldots,99$ and $i = 100,\ldots,999$, respectively. Thus, by comparing the observed frequencies of the digits of the numbers in a data set of numerical records with the expected ones according to the law, one can obtain insights about the authenticity of the data set. Indeed, when a divergence is found, it may be caused by fraudulent behavior of economic agents, namely through the omission of some records from the original data set.

As no information is recorded, data evasion is harder to detect and prove than other data manipulations. In this paper, we apply Benford's law to deal with the problem of detecting and measuring data evasion. In concrete, the paper proposes a procedure to estimate the number of records that were evaded from a data set, which is a new approach in the literature.

The paper is organized as follows. In Section 2, a baseline definition of authenticity of a data set is presented, and in Section 3, the procedure to estimate the number of omitted records from a data set is developed. Section 4 is devoted to the illustration of the application of the proposed procedure and to the assessment of its effectiveness, through an empirical experiment. Section 5 concludes the paper.

2. Benford's law and the authenticity of a data set

To assess the conformity of a data set with Benford's law, several conformity tests and test statistics can be used. For example, Nigrini and Mittermaier (1997) propose the first digits test, the second digits test, the first-two digits test, the number duplication test, the rounding test and the last-two digits test. As test statistics for each conformity test, two of the most common are the Chi-square statistic, which evaluates the global conformity of the data under a conformity test, and the individual standard normally distributed Z-statistic, which assesses the conformity of each particular digit(s) within a conformity test.

With many alternatives for analyzing the conformity of a data set with Benford's law, some assumptions must be made regarding the definition of authenticity of a data set. Once having such a definition, we get a reference that allows to evaluate a given observed data set and argue whether it is authentic. If a data set fails the definition, we assume that some records must have been omitted from the original authentic data set and that it is possible to recover information about the records that were omitted, as it will be explained below. When a record is omitted from a data set, there are two effects: a direct effect decreasing the frequencies of the digits observed by that record, and an indirect effect that increases the observed relative frequencies of all other digits. When intentionally omitting records, it is not reasonable to control these effects given the crossed interaction, which makes it difficult to obtain an artificially conforming data set for the remaining records.

The omission of records in a data set may thus lead the set of remaining records to diverge sufficiently from the law. In particular, it can cause some digits to have significantly lower frequencies than expected, which makes the respective digits good guesses for the digits of the omitted records. Restoring compatibility with the law is thus possible by adding a convenient number of records with the digits for which the frequencies are below the expected. To know how many records to add in each of those frequencies is however more difficult the more conformity tests are involved in the definition of authenticity. This can be seen, for example, if we use the first digits test together with the first-two digits test. In this case, the estimated number of omitted records with first digit 9 should be consistent with the estimated number of omitted records with first-two digits starting with 9.

For simplicity, we avoid the interdependencies between the tests by selecting only the first-two digits test. We select the first-two digits test given that it is the one that captures the most of Benford law's logarithmic nature. Drake and Nigrini (2000) recommend it (or alternatively the first-three digits test if the data set is over 10 000 records) to look for the spikes that signal the digits of the numbers that are more probable to be erroneous, while suggesting that the first digits and the second digits tests must be used only as preliminary conformity tests.

Hereafter, we thus consider the following definition of statistical authenticity of a data set.

2.1 Assumption 1

A data set of T records with T_i records with first-two digits i is statistically authentic if it is conforming with respect to all Z-statistics for the first-two digits test, i.e. if $|T_i/T - e_i|/\sqrt{e_i(1-e_i)/T} \leq Z^*(\alpha)$, $i = 10, \ldots, 99$, where $Z^*(\alpha)$ is the critical value for the Z-statistics at a significance level of α.

The required degree of conformity with Benford's law in Assumption 1 can be adjusted with parameter α. A higher α means being more demanding when validating the conformity of a given observed data set (small ranges for the intervals in which the relative frequencies are conforming). Figure 1 illustrates the behavior of Benford probabilities for each first-two digits combination (from i = 10 to i = 99) and both the

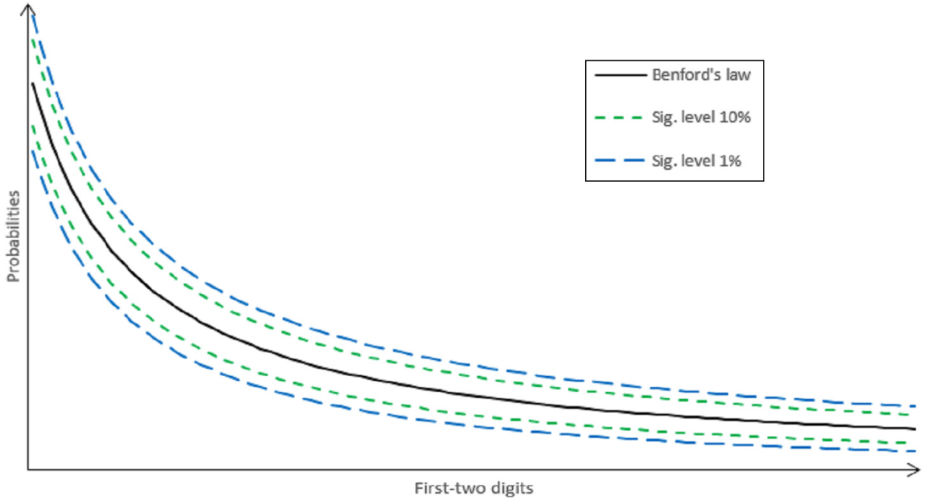

Figure 1.
Benford's law and conformity bounds for $\alpha = 1$ per cent and $\alpha = 10$ per cent

upper and lower bounds for two different conformity levels of the relative frequencies of a given data set ($\alpha = 1$ per cent and $\alpha = 10$ per cent).

Under Assumption 1, the distortions caused by the omission of records are statistically significant, and thus detectable, if there is at least one observed relative frequency outside its bounds.

3. Assessing the number of omitted records

The objective of this section is to estimate the numbers of records that must be added to the observed data set so as to restore conformity according to Assumption 1.

Our approach is based on a mathematical programming model, which results from what follows.

Let n_i be the number of records with first-two digits i in the observed data set, and x_i the number of records with first-two digits i to add to the set of observed records. The total number of records to be added is $k = \Sigma_{i=10}^{99} x_i$[equation (1) in the model below]. The aim of the addition of records is that the new data set satisfies Assumption 1, i.e. $|(n_i + x_i)/(n + k) - e_i|/\sqrt{e_i(1 - e_i)/(n + k)} \leq Z^*(\alpha)$, i = 10,...,99 [equations (3) in the model].

As a fitness measure for the proximity between the frequencies of the new data set and the expected ones according to Benford's law, which are the best guesses for the frequencies of the original authentic data set, we use the Chi-square statistic, given by $(n + k)\,\Sigma_{i=10}^{99}((n_i + x_i)/(n + k) - e_i)^2/e_i$. This statistic constitutes the objective function of the model.

One feature of this function is that it can always be decreased toward zero if k is continuously increased. Of course, in real context, the number of omitted records is limited (a small business is likely to have a lower number of omitted records than a large business). Hence, k must be bounded [$k \leq k^+$, equation (2) of the model].

In summary, the model is as follows:

$$Min\, z = (n + k) \sum_{i=10}^{99} \frac{\left(\frac{n_i + x_i}{n + k} - e_i\right)^2}{e_i}$$

s.t.:

$$\mathrm{k} = \sum_{i=10}^{99} x_i \tag{1}$$

$$\mathrm{k} \leq \mathrm{k}^+ \tag{2}$$

$$\left|\frac{n_i + x_i}{n + k} - e_i\right| / \sqrt{e_i(1 - e_i)/(n + k)} \leq Z^*(\alpha),\ \mathrm{i} = 10,\ldots,99 \tag{3}$$

$$\mathrm{x_i} \geq 0,\ \text{integer} \tag{4}$$

Given that the optimal solution of the model (denoted by z*) is non-increasing with k^+, the choice of k^+ determines the solution. An important issue is thus how to define k^+. A lower bound for k^+ can be k^{lower}, the optimal value of the objective function of an auxiliary problem, given by Min w = k s.t.: (1), (3), (4). This value is the lowest number of records that must be added to the observed data set so that the new data is conforming. An upper bound for k^+, k^{upper}, higher than k^{lower}, may be a context guess for the maximum physically possible number of omitted records. To assist the user of the model in defining k^+ within these bounds, note that after a certain value of k^+, significant improvements in z can be made only at the expense of large increases in k^+. In addition, as the authentic data set may not follow exactly Benford's law (i.e. the relative frequencies of the authentic data set may not be exactly e_i), setting a large k^+ may not be the best practice. We thus increase k^+ from the lower bound toward the upper bound until we observe a relatively small improvement in z*, avoiding an overfitting of the new data set.

In detail, we suggest the following procedure to estimate the number of omitted records (NMR). First, solve the main model under $k^+ = k^{lower}$ so as to obtain the optimal values x_i*, k* and z*. Second, iteratively solve the model for larger values of k^+ until a stopping condition is achieved, which is given by a sufficiently small improvement in the value of z or by k^+ reaching k^{upper}. By adopting this procedure, k^{ref} is the estimate for the total number of omitted records:

```
Procedure NMR
  1. solve the main model considering $k^+ = k^{lower}$ and obtain z* and k*, and let $k^{ref} = k^*$;
  2. solve the main model considering $k^+ = (k^* + k^{upper})/2$ and obtain z** and k**;
  3. while $(z^* - z^{**})/(k^{**} - k^*) > 0.01$ Do
    Begin
      Let $k^+ = (k^{**} + k^{upper})/2$
      Let $z^* = z^{**}$, $k^* = k^{**}$ and $k^{ref} = k^*$
      Solve the main model with $k^+$ and obtain z** and k**
    End
```

4. Empirical experiment

In this section, we conduct an empirical experiment so as to illustrate the application of procedure NMR, considering two different degrees of conformity of the authentic data set with Benford's law. We thus apply the procedure over two different simulated authentic data sets, one satisfying Assumption 1 at a significance level of 10 per cent (strong conformity), named data set 1, and the other satisfying Assumption 1 only at a significance level of 1 per cent (weak conformity), named data set 2. Both data sets consist of 5,000 simulated numerical records, with the absolute frequency of first-two digits i belonging to the interval $[e_iN - N.Z^*(\alpha).\sqrt{e_i(1-e_i)/N};\ e_iN + N.Z^*(\alpha).\sqrt{e_i(1-e_i)/N}]$, where $N = 5{,}000$ and $\alpha = 10$ per cent and $\alpha = 1$ per cent, respectively, for data sets 1 and 2 (see Figures 2 and 3). These sets of records, equivalent to real authentic data sets, constitute the baseline that allows to evaluate the quality of the solutions obtained by applying procedure NMR.

The data sets were then distorted by omitting some of their records. In detail, for each first-two digits combination i, we omitted a random percentage (uniformly between 0 and 100 per cent) of its records. The total number of omitted records was 1,692 in data set 1 and 2,423 in data set 2, so that the distorted data sets are constituted by 3,308 and 2,577 records, respectively (see Figures 2 and 3). Both distorted data sets do not satisfy Assumption 1 at a significance level of 1 per cent.

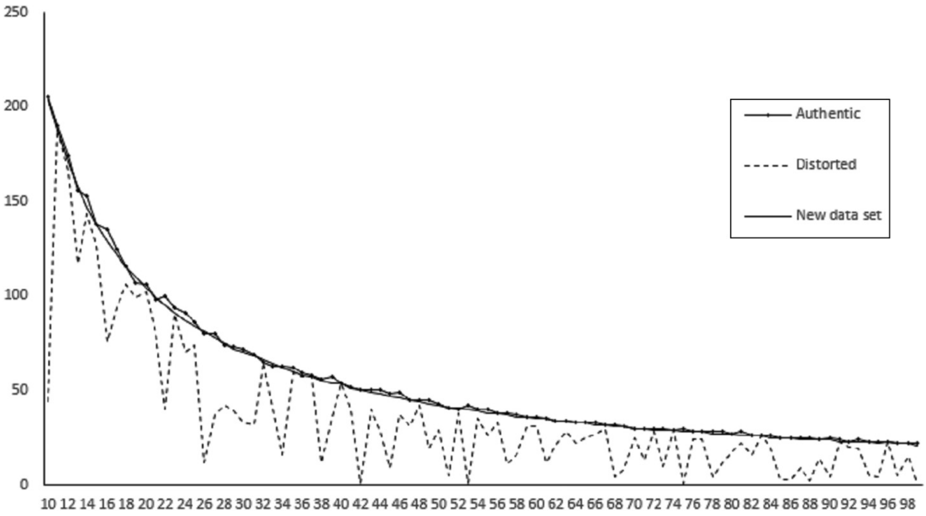

Figure 2.
Results for data set 1

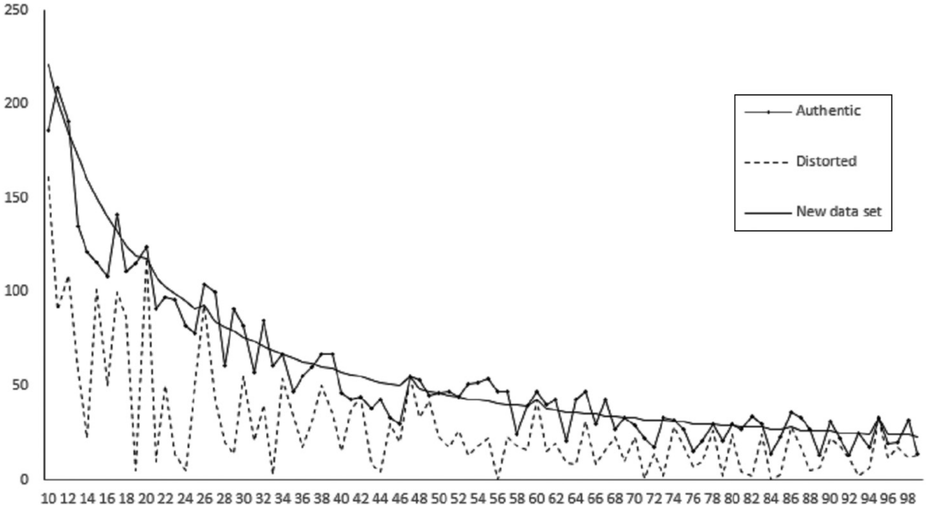

Figure 3.
Results for data set 2

Using $\alpha = 1$ per cent, the lower bounds for k^+ are $k^{lower} = 831$ and $k^{lower} = 2{,}080$, respectively, for the cases of data set 1 and data set 2[1].

In the case of data set 1, k^{upper} was arbitrarily set equal to 2,500, while in the case of data set 2 it was set equal to 3,500.

Table I displays the results obtained when applying procedure NMR on data set 1. The table displays the values of k^+, k^*, z^*, the relative gain in z^*, and a measure of the quality of the solutions (Errors), evaluated by the sum, for all $i = 10,\ldots, 99$, of the absolute difference between x_i^* and the respective true number of omitted records.

As it can be seen in the table, the procedure estimates that 1,612 records were omitted (4.7 per cent below the true number of omitted records, 1,692). Comparing the new data set with the authentic data set, one finds 102 errors (6 per cent of the true number of omitted records). The absolute frequencies for the new data set are displayed in Figure 2. One can observe that the new data set describes well the frequencies of the authentic data set.

Table II displays the results obtained when applying procedure NMR on data set 2. In this case, the procedure estimates that 2,790 records were omitted (15.1 per cent above the true number of omitted records, 2,423), implying a total of 829 errors (34.2 per cent of the true number of omitted records). The absolute frequencies of the new data set are displayed in Figure 3. The fitting is not as good as in the previous case so that the effectiveness of procedure NMR is decreased in this case, particularly with respect to the number of errors. This is because data set 2, even though authentic, does not follow so closely Benford's law as data set 1. This reveals the danger of overfitting the data. The effectiveness of procedure NMR thus seems to increase with the degree of conformity of the authentic data set.

One could also be interested in evaluating the sensibility of the results to the parameter α in the model. However, the effect of a change in α is straightforward. In detail, for example, if α is decreased, the constraints (3) get less restrictive. Therefore, when solving the auxiliary model, one gets a smaller $\mathrm{k}^{\mathrm{lower}}$. This implies that k^{+} can vary within a larger interval. Nevertheless, as procedure NMR promotes continuous increases in k^{+}, the solution of the model converges to stronger levels of conformity, compatible with higher α's. Hence, there is no reason to expect significant changes in the effectiveness of the results by increasing α.

5. Conclusions

The problem of having some economic agents omitting relevant financial, accounting or business information harms the economy as a whole. In this paper, we proposed a procedure that can be used to give insights about the records that were omitted from a data set.

Even though the proposed procedure gives only patterns for the first-two digits distribution, while not properly identifying the specific omitted records, the information provided can be useful to guide auditors in their investigations.

The methodology relies on two key assumptions. First, that there is no other data manipulation than evasion distinguishing the authentic and the observed data sets. Second, that an authentic data set follows Benford's law, which is plausible for most of economic, financial, accounting and business data. As shown in the empirical experiments, the effectiveness of the methodology seems to increase with the degree of conformity of the true data set with the law.

Table I.
Results for data set 1

k^+	k^*	z^*	$(z^* - z^{**})/(k^{**} - k^*)$	Errors
831	831	53.413	–	861
1,612	1,612	0.273	0.068	102
2,056	1,934	0.146	0.000	242

Table II.
Results for data set 2

k^+	k^*	z^*	$(z^* - z^{**})/(k^{**} - k^*)$	Errors
2,080	2,080	22.924	–	787
2,790	2,790	5.092	0.025	829
3,145	3,144	2.276	0.008	891

Concerning its practical applications, procedure NMR can be useful in contexts where the above assumptions are plausible, as for example in detecting unrecorded sales.

As for future research, it could be interesting to investigate reasonable assumptions that would allow to deal with data evasion together with other types of data manipulation simultaneously and introducing some randomness concerning the expected authentic frequencies (alternatively to the consideration of deterministic e_i). This last issue could improve the effectiveness of procedure NMR in cases where the authentic data sets are weakly conforming.

Note

1. Throughout the paper, all the models are solved using the mixed integer nonlinear constrained optimization solvers from the NEOS platform (www.neos-server.org).

References

Benford, F. (1938), “The law of anomalous numbers”, *Proceedings of the American Philosophical Society*, Vol. 78 No. 4, pp. 551-572.

Carlslaw, C. (1988), “Anomalies in income numbers: evidence of goal oriented behavior”, *The Accounting Review*, Vol. LXIII No. 2, pp. 321-327.

Coderre, D. (2009), *Fraud Analysis Techniques using ACL*, John Wiley and Sons, Hoboken, NJ.

Diekmann, A. (2007), “Not the first digit! Using Benford’s Law to detect fraudulent scientific data”, *Journal of Applied Statistics*, Vol. 34 No. 3, pp. 321-329.

Drake, P. and Nigrini, M. (2000), “Computer assisted analytical procedures using Benford’s law”, *Journal of Accounting Education*, Vol. 18 No. 2, pp. 127-146.

Durtschi, C., Hillison, W. and Pacini, C. (2004), “The effective use of Benford’s law to assist in detecting fraud in accounting data”, *Journal of Forensic Accounting*, Vol. 5 No. 1, pp. 17-34.

Gomes da Silva, C. and Carreira, P. (2013), “Selecting audit samples using Benford’s Law”, *Auditing: A Journal of Practice & Theory*, Vol. 32 No. 2, pp. 53-65, doi: 10.2308/ajpt-50340.

Hill, T. (1995), “A statistical derivation of the significant-digit law”, *Statistical Science*, Vol. 10 No. 4, pp. 354-363.

Newcomb, S. (1881), “Note of the frequency of use of the different digits in natural numbers”, *American Journal of Mathematics*, Vol. 4 No. 1, pp. 39-40.

Nigrini, M. (2000), *Continuous Auditing*, Working paper, Ernst & Young Center for Auditing Research and Advanced Technology, University of Kansas, Lawrence, KS.

Nigrini, M. and Mittermaier, L. (1997), “The use of Benford’s law as an aid in analytical procedures”, *Auditing: A Journal of Practice & Theory*, Vol. 16 No. 2, pp. 52-67.

Watrin, C., Struffert, R. and Ullmann, R. (2008), “Benford’s Law: an instrument for selecting tax audit targets?”, *Review of Management Science*, Vol. 2 No. 3, pp. 219-237.

Corresponding author
Pedro Carreira can be contacted at: pedro.carreira@ipleiria.pt