DATA VALIDITY AND STATISTICAL CONFORMITY WITH BENFORD'S LAW

ROY CERQUETI AND MARIO MAGGI

Abstract. Benford's Law is a statistical regularity of a large number of datasets; assessing the compliance of a large dataset with the Benford's Law is a theme of remarkable relevance, mainly for its practical consequences. Such a task can be faced by introducing a statistical distance concept between the empirical distribution of the data and the random variable associated with Benford's Law. This paper deals with the problem of measuring the compliance of a random variable – which can be seen as describing the empirical distribution of a collection of data – with the Benford's Law. It proposes a statistical methodology for detecting the critical values related to conformity/nonconformity with Benford's Law in some well-established cases of statistical distance. The followed approach is grounded on the proper selection of a family of parametric random variables – the lognormal distribution, in our case – and of a reference statistical distance concept – mean absolute deviation. A discussion of the obtained results is carried out on the ground of the existing literature. Moreover, some open problems are also presented.

Keywords: Data science; Benford's Law, Lognormal distribution, Statistical distance.

1. Introduction

In the current debate on data science and its application, a relevant role is played by the assessment of specific statistical regularities of the explored samples. In this context, one has to mention the so-called Benford's Law (BL, hereafter), which has been observed in [31] and formalized in [5]. BL states that the first non zero digits of the numbers contained in a large dataset follow a peculiar distribution, giving higher frequencies to smaller digits. Other versions of BL can be presented for the second digit and the first two digits (and also more than this), but they are out of the scopes of the present paper. For a detailed view of the BL, we refer to some very relevant monographs on the matter (see [7, 21, 25, 34]). For a mathematical treatment of such a law, see e.g. [6, 8, 19, 20].

The popularity of BL – recently acknowledged in a formal way by [29] – is grounded on two main roots. By one side, BL is found to be valid in an incredibly large number of situations and from various scientific contexts like finance (see e.g. [1, 10, 12, 23, 28, 38]), economics (see e.g. [3, 30, 35, 42]), accounting (see e.g. [16, 33]), geophysics and hydrology (see e.g. [2, 13, 39]), electoral studies (see e.g. [11, 24, 36]) and other relevant fields in social science (see e.g. [4, 26, 27]). Under this perspective, we point out that the original contribution in [5] itself contains the analysis of more than 20 thousand data coming from twenty different contexts, and highlights the presence of BL in the most part of the considered datasets; by the other side, BL allows to carry out very intuitive conclusions on the dataset, mainly when it is not valid. Indeed, the non-compliance of a set of numbers with BL calls for a more detailed exploration of such a set. In particular, it is reasonably assumed that the standard case is the validity of BL, so that being out from the BL may suggest that data have been manipulated. Of course, this statement is true under some conditions for the considered data, like having a range of variation with a uniform distribution of several orders of magnitude, being the outcome of several random processes with different probability distributions or being the outcome of a large number of processes of multiplicative type (see e.g. [18]). Obviously, there are no scientific reasons to claim a one-to-one relationship between data manipulation and lack of compliance with BL. However, one can see the invalidity of BL as an occurrence which deserves further research. Thus, BL is often associated with fraud detection and data manipulation (see, e.g. [21] and also the papers quoted above).

In the described framework, a task of crucial relevance is the statistical methodology to be used for assessing the compliance of a statistical variable with BL. By introducing the empirical distribution associated to the considered statistical variable and the random variable related to BL, one can face such a task by measuring the statistical distance between two random variables. There are many classical statistical distance concepts that can be used to this scope. The most popular ones are chi-square, Mean Absolute Deviation (MAD) and Sum of Squared Deviations (SSD). The compliance of a given random variable with BL is statistically proved when the value of the considered statistical distance is below a critical threshold. Needless to say, the value of the critical threshold depends on the employed concept of statistical distance.

In [34], there is a long discussion on statistical distances and critical thresholds. The author states that MAD should be preferred to the chi-square test. Moreover, [34] updates the results obtained by [15] and identifies some ranges of variation of the value of MAD to have close/acceptable/marginally acceptable conformity with BL and nonconformity with BL. The distinction of four cases leads to three critical thresholds (see [34], Table 7.1). Analogously, [21] identifies some critical thresholds of SSD for having perfect/acceptable/marginal conformity and nonconformity with BL. The argument is similar to the one in [34]: MAD and SSD have the common feature to measure the statistical distance between two distributions, with no regard to the sample size.

Unfortunately, there is not an analytical way to derive the critical values associated to MAD and SSD, and the empirical experiments performed by Kossovsky and Nigrini – along with their wisdom and outstanding role in the scientific environment surrounding BL – allow to take for good the proposed critical thresholds (see e.g. [41]). It would be interesting to assess the consistency between the identified critical thresholds for MAD and SSD so that the same dataset has the same level of compliance with BL when evaluated through MAD or SSD. However, to the best of our knowledge, a reliable and rigorous comparison between the critical values of the statistical distances mentioned above is still missing.

This paper fills this gap. Indeed, in accord with [34], we here set MAD as reference statistical distance concept and explore the behavior of chi-square statistical distance and SSD with respect to MAD. Taking MAD as reference distance is a good choice also for a mathematical reason. In fact, the value of MAD between two random variables is bounded, with lower bound equals to zero and upper bound that can be easily computed for the BL (see Section 2).

Furthermore, MAD, SSD and chi-square statistical distance share the same perspective of aggregating the differences – in absolute value or quadratic – between the probabilities, without modifying them through functional transformations, like in the logarithmic case of the Kullback-Leibler divergence. This property allows a direct comparison among their critical values.

The followed approach is based on the computation of MAD between BL and some random variables which depend on a parameter, by letting the parameter change. The selection of the family of parametric random variables is implemented in a proper way so that assigning several values to the parameters leads to values of MAD giving a fine discretization of its range of variation. Then, chi-square and SSD are computed over the same set of random variables, so that a direct comparison between their values and the corresponding value of MAD can be carried out.

In the random variables selection phase, we provide some convincing arguments on the suitability of the lognormal distribution $\log\mathrm{N}(\mu, \sigma)$ with a fixed μ and taking σ as the varying parameter.

Results give an effective device for translating the critical values of MAD in the critical values of chi-square and SSD. Of course, the general idea behind our statistical methodology can be applied to any statistical distance concept and also to other reference random variables – not necessarily BL.

The rest of the paper is organized as follows. Section 2 provides the statistical framework we deal with, with the main preliminaries and notation. Section 3 outlines the methodology used for carrying out the analysis. Section 4 is devoted to the illustration of the results and also gives a discussion of them. Section 5 offers some conclusive remarks and introduces some open problems. Appendix A reports the comparison table between the considered statistical distance measures.

2. The statistical framework

Consider a random variable X_{BL} with nonnegative support such that

$$\delta(d) := P\left(f_{BL}(d)\right) = \log_{10}\left(1 + \frac{1}{d}\right), \qquad \forall d = 1, \dots, 9, \tag{1}$$

where $f_{BL}(d)$ is the event $\{\text{first digit of } X_{BL} = d\}$.

The components of the vector $\delta := (\delta(1), \dots, \delta(9))$ form a probability distribution over the set $\{1, \dots, 9\}$, according to formula (1). Such a distribution is called *Benford's Law* (*BL*, hereafter), and the random variable X_{BL} is the *Benford's variable*. Of course, there exist infinite Benford's variables. Hereafter, we will refer to X_{BL} as one of the elements of the set of random variables satisfying (1).

Let us now consider another random variable Y with nonnegative support and denote by $f_Y(d)$ the event $\{\text{first digit of } Y = d\}$. Define $\gamma_Y(d) := P\left(f_Y(d)\right)$ and the vector $\gamma_Y := (\gamma_Y(1), \dots, \gamma_Y(9))$. Also in this case, γ_Y is a probability distribution over $\{1, \dots, 9\}$.

We define the *difformity* of a random variable Y with respect to the Benford's variable X_{BL} as the statistical distance between the distributions of their first digits, namely δ and γ_Y. The employed concept of distance rules the conceptualization of difformity.

In the context we deal with, we will use the *Mean Absolute Deviation* (*MAD*, hereafter) as the reference difformity measure. For a given random variable Y the MAD with respect to the Benford's variable is defined as the arithmetic mean of the component-wise deviations between δ and γ_Y:

$$\mathrm{MAD}(X_{BL}, Y) = \frac{1}{9}\sum_{d=1}^{9} |\delta(d) - \gamma_Y(d)| \,. \tag{2}$$

MAD is a common measure and test statistics in BL literature. In [15], the authors propose a method and the corresponding MAD empirical critical thresholds to assess the compliance of data with BL. Moreover, in [34] the thresholds are updated, and the justification of the use of MAD is clarified: tests based on MAD thresholds do not suffer from the high reject ratio produced by other standard tests in case of large samples.

Therefore, in this specific case, the following result holds true.

Proposition 1. *The random variable $\bar{Y}$ with the maximum difformity with respect to the Benford's variable is such that*

$$\gamma_{\bar{Y}}(d) = \begin{cases} 0 & \textit{for } d = 1, \dots, 8 \\ 1 & \textit{for } d = 9 \end{cases} \tag{3}$$

Proof. The distribution $\bar{Y}$ which maximizes the MAD with respect to the BL solves the optimization problem

$$\max_{\gamma(d), d=1,\dots,9} \sum_{d=1}^{9} |\gamma(d) - \delta(d)| \tag{4}$$

with the constraints

$$\begin{cases} \sum\limits_{d=1}^{9} \gamma(d) = 1 \\ \gamma(d) \geq 0, d = 1, \dots, 9 \end{cases} \tag{5}$$

The problem (4-5) can be easily linearized; therefore the solution lays on the vertexes of the admissible set (or in a convex combinations of the vertexes). The vertex of the admissible set are such that one element of γ is equal to 1, whereas all the other are null. Let k be index of the unit component, the MAD on the vertex k is

$$\sum_{d \neq k} \delta(d) + 1 - \delta(k) = 1 - \delta(k) + 1 - \delta(k) = 2(1 - \delta(k))$$

which is maximized for the smallest value of $\delta(d)$, i.e. $\delta(9)$, therefore, the k which maximizes the MAD is 9, leading to define the random variable $\bar{Y}$ with the maximum difformity with respect to the Benford's variable as the one satisfying (3). □

There is an infinite number of random variables satisfying (3), like for example $U(9, 10)$; $U(9 \times 10^k, 10^{k+1})$, $\forall k \in \mathbb{Z}$; the Dirac mass (see [14]) $\delta_{\mathrm{Dirac}}(x - \gamma)$, with $\gamma \in (9 \times 10^k, 10^{k+1})$, $\forall k \in \mathbb{Z}$. Thus, we will refer hereafter to $\bar{Y}$ as one of the elements of the set of the random variables satisfying (3).

The definition of the random variable $\bar{Y}$ through (3) leads to the upper bound for the MAD between the random variables X_{BL} and $\bar{Y}$. Specifically,b

$$\mathrm{MAD}(X_{BL},\bar{Y}) = \frac{1}{9}\sum_{d=1}^{9}|\delta(d)-\gamma_{\bar{Y}}(d)| = \frac{2}{9}(1-\delta(9)) \approx 0.21. \tag{6}$$

Thus, for each random variable Y, formula (6) assures that $\mathrm{MAD}(X_{BL},Y) \in [0,0.21]$, where the corner cases are

$$\mathrm{MAD}(X_{BL},Y)=0 \text{ for } Y=X_{BL}; \qquad \mathrm{MAD}(X_{BL},Y)=0.21 \text{ for } Y=\bar{Y}.$$

MAD will be compared with other well-established statistical distance measures, which are often used for assessing the compliance of a random variable with the BL. In particular, we consider the following ones.

- *Sum of Squared Deviations* (*SSD*):

$$\mathrm{SSD}(X_{BL},Y) = \sum_{d=1}^{9}(\delta(d)-\gamma_Y(d))^2; \tag{7}$$

- χ^2:

$$\chi^2(X_{BL},Y) = \sum_{d=1}^{9}\frac{(\delta(d)-\gamma_Y(d))^2}{\delta(d)}. \tag{8}$$

Interestingly, Eq. (8) presents the χ^2 "distance" or "deviation" or "dissimilarity", i.e. a measure of the difference between two distributions. Such a distance is also denoted as Neyman χ^2 distance (see e.g. [9]), and it is widely used for applications (see e.g. [22, 32]). Also the selection of the χ^2 as in (8) meets the requirement of measuring the dissimilarity between distributions, without being affected by the sample size. This is consistent with the arguments provided by [34] and avoids the common issue related to test statistics, which tend to be too powerful with large samples, yielding to reject the BL too often.

3. Methodology

This section outlines the methodology employed for exploring the relationship between the MAD and the other statistical distance measures used for assessing the compliance with the BL.

The followed approach can be described in a step-wise form.

- We select a family of random variables with nonnegative support and dependent on some parameters – we here take only one real parameter α. We denote the generic random variables with parameter α by $Y(\alpha)$. The probability

$\gamma_{Y(\alpha)}(d), d = 1, \dots, 9$ can be computed through the bilateral series

$$\gamma_{Y(\alpha)}(d) = \sum_{h=-\infty}^{+\infty} P\left[d \cdot 10^h \leq Y(\alpha) < (d+1) \cdot 10^h\right] =$$

$$= \sum_{h=-\infty}^{+\infty} \left[F_{Y(\alpha)}\left((d+1) \cdot 10^h\right) - F_{Y(\alpha)}\left(d \cdot 10^h\right)\right], \tag{9}$$

where $F_{Y(\alpha)} : \mathbb{R} \to [0, 1]$ is the cumulative distribution function of the random variable $Y(\alpha)$.
Therefore, the probability distribution γ of the first digit can be numerically approximated through (9) with great precision (see below).
The selection of such a family is implemented in a proper way (see the next step and the details in the next section).

- We compute the MAD between $Y(\alpha)$ and X_{BL} by letting the value of α vary. Specifically, we consider $\alpha = \alpha_1, \dots, \alpha_n$ and denote $\text{MAD}_k := \text{MAD}(X_{BL}, Y(\alpha_k))$, for each $k = 1, \dots, n$. The selection of the family of random variable $(Y(\alpha) : \alpha \in \mathbb{R})$ and of the set $\{\alpha_1, \dots, \alpha_n\}$ are such that the discrete set $\{\text{MAD}_1, \dots, \text{MAD}_n\}$ represents a fine discretization of the range of variation of the MAD, i.e. $[0, 0.21]$. As we will see in the next section, we consider the lognormal random variables $Y \sim \text{logN}(\mu, \sigma)$, by conveniently fixing the value of μ and letting σ play the role of parameter α.
- We compute $\text{SSD}_k := \text{SSD}(X_{BL}, Y(\alpha_k))$ and $\chi^2_k := \chi^2(X_{BL}, Y(\alpha_k))$ according to formulas (7) and (8), respectively, for each $k = 1, \dots, n$.
- We compare SSD_k and χ_k with MAD_k for each $k = 1, \dots, n$. In so doing, we derive the relationship between the values of the MAD's and the ones of the other considered statistical distance measures.

The bilateral sum in (9) is truncated from $-H$ to H, where H is increasing with the parameters σ and μ, to account for dispersion and magnitude. For the results that will be presented in the next section, we set

$$H = \lceil |\log_{10}\left(e^{\sigma + \frac{1}{2}\mu}\right)| \rceil + 3,$$

where $\lceil \cdot \rceil$ is the *ceil* function, and the additional 3 widens the range by 3 orders of magnitudes (this helps in assuring to span the relevant range, with a small additional computational cost). From many preliminary numerical trials, this bounds have proven to provide a high accuracy and a good trade-off between computational speed and precision; the probability left outside these bounds is negligible.

4. Results and discussion

As already preannounced in the previous section, we identify the family of the random variables to be used in the analysis. We select the lognormal distribution $\text{logN}(\mu, \sigma)$ with a fixed value of μ and by letting the σ vary. It is well known that the lognormal distribution gets closer to BL as the parameter σ increases, as shown by simulation

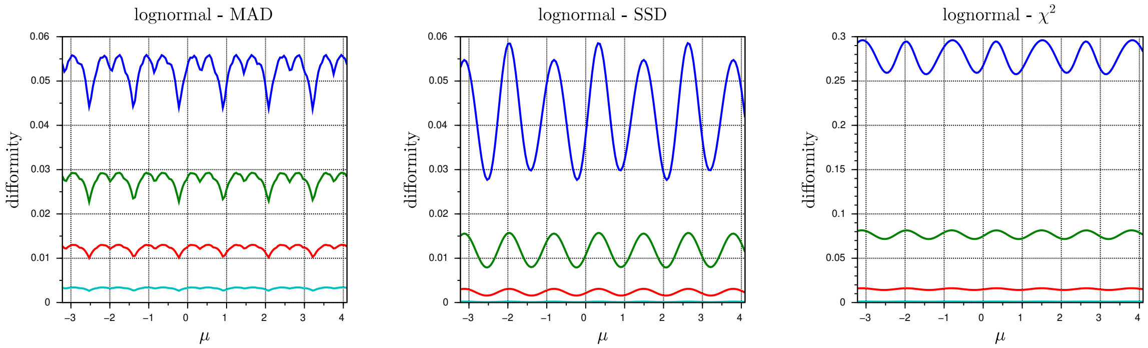

Figure 1. Difformity values for the lognormal $\log\mathrm{N}(\mu, \sigma)$, for $\sigma = 0.5$ (blue), $\sigma = 0.65$ (green), $\sigma = 0.8$ (red), $\sigma = 1$ (cyan).

studies (see, e.g. [17] and [37]). To further support such a choice, we present some preliminary investigations of the lognormal random variable in our context.

At this aim, we tackle the problem by analyzing how changing the two parameters affects the statistical distance between the lognormal random variable and the Benford's one.

In this respect, the parameter μ seems not to play a relevant role in determining the values of the considered statistical distance measures. Indeed, we notice the weak effect of the variation in μ highlighted in Table 1.

Table 1. Values of MAD between the lognormal random variable and the Benford's variable for different values of the parameters μ and σ.

$\mu \setminus \sigma$	0.1	0.2	0.4	0.8	1.6
0.5	0.149391	0.116816	0.077774	0.013022	0.000010
1	0.155075	0.135506	0.070392	0.011353	0.000009
2	0.173548	0.135200	0.069979	0.011294	0.000009
4	0.166567	0.135780	0.077769	0.013022	0.000010
8	0.154955	0.134495	0.076634	0.012235	0.001562

Such an argument is further supported by Figure 1, where the effect of μ on the statistical distances for various values of the parameter σ is shown. Note that all the statistical distances – and not only MAD – remain bounded on a band for a given value of σ. Observe that a modification of μ is equivalent to a rescaling of the variable so that increasing μ by a quantity – say $\theta \in \mathbb{R}$ – is equivalent to multiplying the random variable by e^θ. Hence, the fact that the statistical distance is not constant with respect to μ – and, therefore, with respect to the scale factor e^θ applied to the lognormal random variable – confirms that the distribution of the first digit of the lognormal is not scale-invariant.

However, the statistical distances are bounded on a very narrow strip depending on the value of the parameter σ. This supports that the key parameter governing the statistical distance of the lognormal from the Benford's variable is σ.

The value of μ for the lognormal random variables is conveniently set to $\mu = \log(9.5)$ while – as already stated above – the parameter σ plays the role of the varying parameter

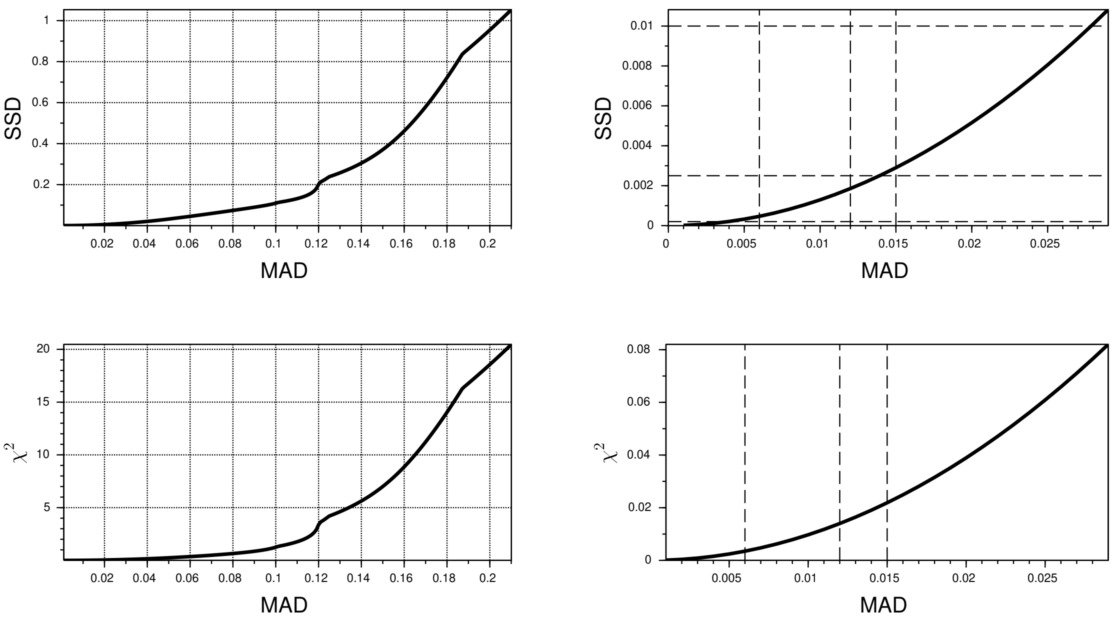

Figure 2. Values of various statistical distances for the corresponding MAD thresholds reportd in Table 2. $\mu = \ln 9.5$. In the right panes a magnified region is displayed. Here the vertical lines correspond to the Nigrini critical values for MAD. In the top right panel, the horizontal lines correspond to the Kossovsky critical values for SSD.

α, so that $\text{logN}(\log(9.5), \sigma) = Y(\alpha)$ by setting $\sigma = \alpha$. Indeed, by selecting $\mu = \log(9.5)$, one observes an increasing similarity as $\sigma \to 0^+$ between $\text{logN}(\log(9.5), \sigma)$ and the Dirac mass assigning unitary probability to $x = 9.5$, which then satisfies (3). Therefore, a proper selection of $\alpha_1, \ldots, \alpha_n$ leads to a fine discretization of $[0, 0.21]$, which is the range of variation of the MAD.

Table 2 in Appendix A reports the results of the analysis, while Figure 2 offers the graphical representation of the relationship between the values of MAD and those of the other considered statistical distances.

Observe that the expertise-driven critical values proposed by Nigrini [34] and Kossovsky [21] can now be compared. Although they seem to be reasonably coupled in both cases, they do not exactly correspond to each other. In fact, while the close conformity threshold by Kossovsky is more severe than the one by Nigrini, the other two thresholds are less strict. So this comparison allows us to conclude that the two ways those authors consider the conformity to BL may depend on their experience and their insights.

Interestingly, one can also observe a non-regular behaviour of the relationship between MAD and both SSD and chi-square, with a small jump around MAD=0.12 and an increasing rate of growth of SSD and chi-square as MAD increases.

5. Conclusions and open questions

This paper provides a methodological framework for a rigorous comparison the critical values of three well-established statistical distances – i.e., MAD, SSD and chi-square – which are commonly used for assessing the compliance with the BL. According to the compliance critical values proposed in [34], we take MAD as reference statistical distance. In comparing MAD with SSD, we also include a discussion on the conformity thresholds introduced by [21].

Our setting and the obtained results open new questions, rather than writing a final word on the problem of the compliance of data to the BL. Indeed, due to the relevant consequences of the findings of nonconformity in forensic activity, the different interpretations of the concept of compliance with the BL can leave room to different conclusions about the data. Our contribution allows to convert the MAD results into SSD ones easily and vice-versa.

Furthermore, our study is associated with a deep analysis of the distribution of the first significant digit of the lognormal random variables. In fact, although some simulated results are available in the reference literature (e.g., see [17]), this paper eliminates the Monte Carlo errors completely by providing a numerical analysis of the first digit distribution (through the computation of (9)). This is a methodological contribution which can be generalized to any distribution with nonnegative support, hence allowing the statistical-probabilistic study of the compliance with BL in a broad set of contexts.

Lastly, the employed methodology can be extended to the distribution of the first k significant digits of the elements of a large enough dataset, with $k = 1, 2, \dots$. Even if this will add complexity to the analysis, it would also allow discussing the critical values for compliance with the BL provided by Nigrini [34] and Kossovsky [21] for the first two and three digits.

References

[1] Ausloos, M., Castellano, R., & Cerqueti, R. (2016). Regularities and discrepancies of credit default swaps: a data science approach through Benford's law. *Chaos, Solitons and Fractals*, 90, 8-17.

[2] Ausloos, M., Cerqueti, R., & Lupi, C. (2017). Long-range properties and data validity for hydrogeological time series: The case of the Paglia river. *Physica A: Statistical Mechanics and its Applications*, 470, 39-50.

[3] Ausloos, M., Cerqueti, R., & Mir, T. A. (2017). Data science for assessing possible tax income manipulation: The case of Italy. *Chaos, Solitons and Fractals*, 104, 238-256.

[4] Ausloos, M., Herteliu, C., & Ileanu, B. (2015). Breakdown of Benford's law for birth data. *Physica A: Statistical Mechanics and its Applications*, 419, 736-745.

[5] Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4), 551-572.

[6] Berger, A., & Hill, T. P. (2011). A basic theory of Benford's Law. *Probability Surveys*, 8, 1-126.

[7] Berger, A., & Hill, T. P. (2015). *An introduction to Benford's law.* Princeton University Press.

[8] Berger, A., & Hill, T. P. (2020). The mathematics of Benford's law: a primer. *Statistical Methods & Applications*, https://doi.org/10.1007/s10260-020-00532-8.

[9] Broniatowski, M., & Leorato, S. (2006). An estimation method for the Neyman chi-square divergence with application to test of hypotheses. *Journal of Multivariate Analysis*, 97(6), 1409-1436.

[10] Clippe, P., & Ausloos, M. (2012). Benford's law and Theil transform of financial data. *Physica A: Statistical Mechanics and its Applications*, 391(24), 6556-6567.

[11] Deckert, J., Myagkov, M., & Ordeshook, P. C. (2011). Benford's Law and the detection of election fraud. *Political Analysis*, 19(3), 245-268.

[12] De Ceuster, M. J., Dhaene, G., & Schatteman, T. (1998). On the hypothesis of psychological barriers in stock markets and Benford's Law. *Journal of Empirical Finance*, 5(3), 263-279.

[13] Díaz, J., Gallart, J., & Ruiz, M. (2015). On the ability of the Benford's Law to detect earthquakes and discriminate seismic signals. *Seismological Research Letters*, 86(1), 192-201.

[14] Dirac, P. A. M. (1981). *The principles of quantum mechanics (No. 27).* Oxford University Press.

[15] Drake, P. D., & Nigrini, M. J. (2000). Computer assisted analytical procedures using Benford's Law. *Journal of Accounting Education*, 18(2), 127-146.

[16] Durtschi, C., Hillison, W., & Pacini, C. (2004). The effective use of Benford's law to assist in detecting fraud in accounting data. *Journal of Forensic Accounting*, 5(1), 17-34.

[17] Fang, G., & Chen, Q. (2020). Several common probability distributions obey Benford's law. *Physica A: Statistical Mechanics and its Applications*, 540, 123129.

[18] Fernandez-Gracia, J., & Lacasa, L. (2018). Bipartisanship breakdown, functional networks, and forensic analysis in Spanish 2015 and 2016 national elections. *Complexity*, Article ID 9684749.

[19] Hill, T. P. (1995). Base-invariance implies Benford's law. *Proceedings of the American Mathematical Society*, 123(3), 887-895.

[20] Hill, T. P. (1995). A statistical derivation of the significant-digit law. *Statistical Science*, 10(4), 354-363.

[21] Kossovsky A. E. (2014). *Benford's Law: Theory, the General Law of Relative Quantities, and Forensic Fraud Detection Applications (Vol. 3).* Singapore: World Scientific.

[22] Leonenko, G., Los, S. O., & North, P. R. (2013). Statistical distances and their applications to biophysical parameter estimation: Information measures, M-estimates, and minimum contrast methods. *Remote Sensing*, 5(3), 1355-1388.

[23] Ley, E. (1996). On the peculiar distribution of the US stock indexes' digits. *The American Statistician*, 50(4), 311-313.

[24] Mebane, W. R. (2011). Comment on "Benford's Law and the detection of election fraud". *Political Analysis*, 19(3), 269-272.

[25] Miller, S. J. (Ed.). (2015). *Benford's Law.* Princeton University Press.

[26] Mir, T. A. (2012). The law of the leading digits and the world religions. *Physica A: Statistical Mechanics and its Applications*, 391(3), 792-798.

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

[28] Mir, T. A. (2016). The leading digit distribution of the worldwide illicit financial flows. *Quality & Quantity*, 50(1), 271-281.

[29] Mir, T. A., & Ausloos, M. (2018). Benford's law: A "sleeping beauty" sleeping in the dirty pages of logarithmic tables. *Journal of the Association for Information Science and Technology*, 69(3), 349-358.

[30] Mir, T. A., Ausloos, M., & Cerqueti, R. (2014). Benford's law predicted digit distribution of aggregated income taxes: the surprising conformity of Italian cities and regions. *The European Physical Journal B*, 87(11), 261.

[31] Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, 4(1), 39-40.

[32] Nielsen, F., & Nock, R. (2013). On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Processing Letters*, 21(1), 10-13.

[33] Nigrini, M. J. (1999). I've got your number. *Journal of Accountancy*, 187(5), 79-83.

[34] Nigrini, M. J. (2012). *Benford's Law: Applications for forensic accounting, auditing, and fraud detection (Vol. 586).* John Wiley & Sons.

[35] Nye, J., & Moul, C. (2007). The political economy of numbers: on the application of Benford's law to international macroeconomic statistics. *The BE Journal of Macroeconomics*, 7(1).

[36] Pericchi, L., & Torres, D. (2011). Quick Anomaly Detection by the Newcomb—Benford Law, with Applications to Electoral Processes Data from the USA, Puerto Rico and Venezuela. *Statistical Science*, 26(4), 502–516.

[37] Posch, P. N. (2008). A survey on sequences and distribution functions satisfying the first-digit-law. *Journal of Statistics and Management Systems*, 11(1), 1–19.

[38] Riccioni, J., & Cerqueti, R. (2018). Regular paths in financial markets: Investigating the Benford's law. *Chaos, Solitons and Fractals*, 107, 186–194.

[39] Sambridge, M., Tkalcic, H., & Jackson, A. (2010). Benford's law in the natural sciences. *Geophysical research letters*, 37(22), L22301.

[40] Shi, J., Ausloos, M., & Zhu, T. (2018). Benford's law first significant digit and distribution distances for testing the reliability of financial reports in developing countries. *Physica A: Statistical Mechanics and its Applications*, 492, 878-888.

[41] Slepkov, A. D., Ironside, K. B., & Di Battista, D. (2015). Benford's Law: Textbook exercises and multiple-choice testbanks. *PLoS One*, 10(2), e0117972.

[42] Todter, K. H. (2009). Benford's Law as an Indicator of Fraud in Economics. *German Economic Review*, 10(3), 339–351.

Appendix A. Comparison table

Table 2: Correspondence table between the MAD and the other statistical distances: SSD and χ^2. Also, the parameter σ of the lognormal distribution is shown. In bold the critical MAD values for Nigrini [34] and the critical SSD values for Kossovsky [21] – i.e., 0.006, 0.012, 0.015 and 0.0002, 0.0025, 0.0101 for close/acceptable/marginally acceptable conformity with BL, respectively. Due to the discretization on MAD, the highest critical value for SSD is indicated as 0.0101, rather than as 0.01 as it is reported in [21].

MAD thresholds	σ	SSD	χ^2
0.001	1.1460	0.0000	0.0001
0.002	1.0617	0.0001	0.0004
0.003	1.0091	0.0001	0.0009
0.004	0.9701	**0.0002**	0.0016
0.005	0.9387	0.0003	0.0024
0.006	0.0018	0.0005	0.0035
0.007	0.8893	0.0006	0.0048
0.008	0.8688	0.0008	0.0062
0.009	0.8505	0.0010	0.0079
0.010	0.8337	0.0013	0.0097
0.011	0.8180	0.0016	0.0118
0.012	0.0072	0.0019	0.0140
0.013	0.7903	0.0022	0.0164
0.014	0.7774	**0.0025**	0.0191
0.015	0.0113	0.0029	0.0219
0.016	0.7541	0.0033	0.0249
0.017	0.7432	0.0037	0.0281
0.018	0.7328	0.0042	0.0315
0.019	0.7227	0.0047	0.0351
0.020	0.7130	0.0052	0.0390
0.021	0.7038	0.0057	0.0429
0.022	0.6947	0.0062	0.0472
0.023	0.6862	0.0068	0.0515
0.024	0.6777	0.0074	0.0561
0.025	0.6696	0.0080	0.0609
0.026	0.6615	0.0087	0.0659
0.027	0.6539	0.0094	0.0710
0.028	0.6463	**0.0101**	0.0764
0.029	0.6389	0.0108	0.0820
0.030	0.6315	0.0116	0.0879

MAD thresholds	σ	SSD	χ^2
0.031	0.6246	0.0124	0.0938
0.032	0.6177	0.0132	0.0999
0.033	0.6107	0.0140	0.1064
0.034	0.6040	0.0149	0.1130
0.035	0.5973	0.0158	0.1199
0.036	0.5909	0.0167	0.1269
0.037	0.5847	0.0176	0.1340
0.038	0.5782	0.0186	0.1417
0.039	0.5722	0.0195	0.1491
0.040	0.5660	0.0205	0.1571
0.041	0.5600	0.0216	0.1651
0.042	0.5540	0.0226	0.1734
0.043	0.5480	0.0237	0.1821
0.044	0.5422	0.0248	0.1907
0.045	0.5364	0.0259	0.1996
0.046	0.5306	0.0271	0.2088
0.047	0.5249	0.0283	0.2184
0.048	0.5191	0.0295	0.2283
0.049	0.5136	0.0307	0.2380
0.050	0.5080	0.0319	0.2481
0.051	0.5025	0.0332	0.2585
0.052	0.4969	0.0345	0.2692
0.053	0.4914	0.0358	0.2801
0.054	0.4859	0.0372	0.2914
0.055	0.4808	0.0384	0.3021
0.056	0.4755	0.0398	0.3135
0.057	0.4704	0.0411	0.3247
0.058	0.4653	0.0424	0.3361
0.059	0.4602	0.0438	0.3478
0.060	0.4554	0.0451	0.3593
0.061	0.4503	0.0465	0.3715
0.062	0.4452	0.0479	0.3841
0.063	0.4404	0.0493	0.3963
0.064	0.4355	0.0506	0.4088
0.065	0.4305	0.0521	0.4221
0.066	0.4256	0.0535	0.4351
0.067	0.4208	0.0549	0.4484
0.068	0.4157	0.0564	0.4625
0.069	0.4108	0.0578	0.4763
0.070	0.4060	0.0592	0.4904
0.071	0.4012	0.0607	0.5047
0.072	0.3961	0.0622	0.5200
0.073	0.3912	0.0636	0.5349

MAD thresholds	σ	SSD	χ^2
0.074	0.3864	0.0651	0.5501
0.075	0.3815	0.0666	0.5656
0.076	0.3765	0.0681	0.5821
0.077	0.3716	0.0695	0.5982
0.078	0.3665	0.0711	0.6154
0.079	0.3617	0.0725	0.6322
0.080	0.3566	0.0740	0.6501
0.081	0.3515	0.0755	0.6684
0.082	0.3465	0.0770	0.6870
0.083	0.3411	0.0786	0.7070
0.084	0.3361	0.0801	0.7265
0.085	0.3308	0.0817	0.7474
0.086	0.3254	0.0832	0.7689
0.087	0.3201	0.0847	0.7909
0.088	0.3146	0.0863	0.8145
0.089	0.3093	0.0879	0.8377
0.090	0.3035	0.0895	0.8637
0.091	0.2977	0.0912	0.8906
0.092	0.2920	0.0928	0.9184
0.093	0.2860	0.0945	0.9484
0.094	0.2797	0.0963	0.9808
0.095	0.2733	0.0982	1.0158
0.096	0.2668	0.1001	1.0525
0.097	0.2599	0.1021	1.0938
0.098	0.2527	0.1043	1.1389
0.099	0.2451	0.1067	1.1900
0.100	0.2370	0.1093	1.2480
0.101	0.2283	0.1123	1.3160
0.102	0.2223	0.1145	1.3661
0.103	0.2177	0.1162	1.4067
0.104	0.2130	0.1181	1.4494
0.105	0.2084	0.1200	1.4942
0.106	0.2038	0.1220	1.5414
0.107	0.1990	0.1242	1.5937
0.108	0.1943	0.1264	1.6463
0.109	0.1895	0.1289	1.7048
0.110	0.1847	0.1315	1.7669
0.111	0.1796	0.1344	1.8363
0.112	0.1745	0.1375	1.9104
0.113	0.1692	0.1410	1.9935
0.114	0.1636	0.1450	2.0871
0.115	0.1579	0.1495	2.1929
0.116	0.1516	0.1548	2.3180

MAD thresholds	σ	SSD	χ^2
0.117	0.1450	0.1612	2.4666
0.118	0.1373	0.1694	2.6567
0.119	0.1281	0.1810	2.9232
0.120	0.1154	0.2009	3.3725
0.121	0.1092	0.2128	3.6380
0.122	0.1064	0.2186	3.7676
0.123	0.1036	0.2249	3.9051
0.124	0.1009	0.2315	4.0514
0.125	0.0981	0.2386	4.2072
0.126	0.0967	0.2424	4.2890
0.127	0.0956	0.2456	4.3592
0.128	0.0942	0.2496	4.4460
0.129	0.0928	0.2537	4.5359
0.130	0.0916	0.2573	4.6132
0.131	0.0902	0.2617	4.7089
0.132	0.0891	0.2656	4.7914
0.133	0.0877	0.2703	4.8936
0.134	0.0866	0.2744	4.9817
0.135	0.0852	0.2795	5.0912
0.136	0.0840	0.2840	5.1856
0.137	0.0829	0.2885	5.2831
0.138	0.0815	0.2942	5.4044
0.139	0.0803	0.2992	5.5092
0.140	0.0792	0.3043	5.6177
0.141	0.0778	0.3107	5.7529
0.142	0.0766	0.3163	5.8701
0.143	0.0755	0.3220	5.9915
0.144	0.0743	0.3280	6.1174
0.145	0.0729	0.3356	6.2748
0.146	0.0718	0.3421	6.4115
0.147	0.0706	0.3489	6.5536
0.148	0.0695	0.3560	6.7013
0.149	0.0683	0.3634	6.8550
0.150	0.0672	0.3712	7.0150
0.151	0.0660	0.3792	7.1817
0.152	0.0651	0.3860	7.3200
0.153	0.0639	0.3947	7.4995
0.154	0.0628	0.4038	7.6868
0.155	0.0616	0.4134	7.8821
0.156	0.0607	0.4213	8.0446
0.157	0.0595	0.4317	8.2557
0.158	0.0586	0.4403	8.4313
0.159	0.0575	0.4515	8.6595

MAD thresholds	σ	SSD	χ^2
0.160	0.0565	0.4609	8.8494
0.161	0.0554	0.4731	9.0964
0.162	0.0545	0.4833	9.3019
0.163	0.0535	0.4938	9.5146
0.164	0.0526	0.5048	9.7350
0.165	0.0517	0.5162	9.9630
0.166	0.0508	0.5279	10.1991
0.167	0.0499	0.5401	10.4433
0.168	0.0489	0.5528	10.6960
0.169	0.0480	0.5659	10.9571
0.170	0.0471	0.5794	11.2270
0.171	0.0462	0.5934	11.5056
0.172	0.0452	0.6079	11.7932
0.173	0.0445	0.6191	12.0147
0.174	0.0436	0.6344	12.3178
0.175	0.0429	0.6462	12.5509
0.176	0.0420	0.6623	12.8694
0.177	0.0413	0.6747	13.1139
0.178	0.0404	0.6916	13.4472
0.179	0.0397	0.7046	13.7024
0.180	0.0388	0.7222	14.0493
0.181	0.0381	0.7357	14.3141
0.182	0.0372	0.7541	14.6729
0.183	0.0365	0.7680	14.9458
0.184	0.0355	0.7868	15.3140
0.185	0.0349	0.8011	15.5928
0.186	0.0342	0.8155	15.8732
0.187	0.0332	0.8348	16.2485
0.188	0.0328	0.8444	16.4363
0.189	0.0323	0.8541	16.6239
0.190	0.0319	0.8637	16.8112
0.191	0.0314	0.8733	16.9977
0.192	0.0312	0.8781	17.0906
0.193	0.0307	0.8876	17.2756
0.194	0.0302	0.8971	17.4592
0.195	0.0298	0.9065	17.6411
0.196	0.0293	0.9158	17.8210
0.197	0.0289	0.9250	17.9986
0.198	0.0284	0.9340	18.1735
0.199	0.0279	0.9429	18.3453
0.200	0.0272	0.9560	18.5966
0.201	0.0268	0.9644	18.7593
0.202	0.0263	0.9726	18.9177

MAD thresholds	σ	SSD	χ^2
0.203	0.0256	0.9845	19.1463
0.204	0.0252	0.9921	19.2923
0.205	0.0245	1.0029	19.5006
0.206	0.0238	1.0131	19.6953
0.207	0.0231	1.0225	19.8753
0.208	0.0222	1.0338	20.0910
0.209	0.0212	1.0435	20.2776
0.210	0.0201	1.0536	20.4689

Roy Cerqueti (Corresponding author) — Sapienza University of Rome, Department of Social and Economic Sciences. P.le A. Moro 5, I-00185, Rome, Italy

and

London South Bank University, School of Business. 103 Borough Road, London, SE1 0AA, United Kingdom
Email address: roy.cerqueti@uniroma1.it

Mario Maggi — University of Pavia, Department of Economics and Management. Via S. Felice 5, I-27100, Pavia, Italy
Email address: mario.maggi@unipv.it