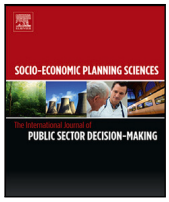
Socio-Economic Planning Sciences 95 (2024) 102008

Contents lists available at ScienceDirect

Socio-Economic Planning Sciences

journal homepage: www.elsevier.com/locate/seps



Validating Benfordness on contaminated data

Marco Di Marzio, Stefania Fensore, Chiara Passamonti *

University of Chieti-Pescara, Viale Pindaro 42, Pescara, 65129, Italy

ARTICLE INFO

Keywords:
Benford
Deconvolution method
Fraud detection
Kernel density estimation

ABSTRACT

Benford's law is a mathematical model, very recurrent in practice for a wide variety of datasets, used to represent the frequencies of digits. A well-established usage of Benfordness statistical testing lies within investigations aimed to ascertain if balance sheet and income statement data are genuine. A typical, frustrating problem of Benfordness statistical tests on big, practical datasets is that they often provide *p-values*smaller than expected when the Benfordness null hypothesis is very realistic. A possible reason is that data are contaminated by some kind of noise. In this paper we propose the deconvolution approach to alleviate this issue, using both simulated and real data.

1. Introduction

Benford's law, also called the *first-digit law*, is a probability distribution for significant digits (the first non-zero digit appearing in the decimal expansion of a number), that characterized several real-world datasets. In 1881 astronomer Newcomb and, more than fifty years later, physicist Benford independently noticed that in a large collection of numbers the frequency distribution of the first significant digit is described by a logarithmic model, and not by a uniform one, as it could be expected (see [1,2]). Specifically, the frequencies of the digits are heavily biased towards lower ones like 1, 2, and 3, over the higher digits, such as 7, 8 and 9. As an example, consider that the frequency of numbers with leading digit 1 is about 30%, while the frequency of the leading digit 9 is more or less 4.6%.

Benford's law has attracted the interest of many researchers in different areas. The reason is that several real-life datasets follow the Benford's law in many fields such as economics (incomes, insurance claims, electricity/telephone bills), finance (loan data, stock and house prices, accounting/credit card transactions, customer balances), social (populations of cities, death rates) and environmental (lengths and flow rates of rivers) sciences and astrophysics (diameter of planets, stellar and galaxy distances). Moreover, this law is used as a tool to check and verify the authenticity of data. In fact, a well-established belief is that deviation from the expected law for genuine observations could indicate possible data manipulation. This idea has been developed in a growing number of fields, involving financial statements (Nigrini [3]), electoral processes (see, for example, [4,5]) and international trade [6,7]. Moreover, recent studies have used Benford's law to assess the accuracy of COVID-19 data reported by different countries (see, for example, [8,9]).

A simple way to investigate the conformity of data to Benford's law consists in calculating a discrepancy index and then judging if it is big (small) enough to informally disprove (support) Benford's law, without relying on distributional arguments (see, for example, [10]). A possible discrepancy index is represented by the mean absolute deviation (MAD) which compares the expected proportion (EP) with the actual one (AP) as $MAD = 1/K \sum_{i=1}^{K} |EP - AP|$, where K is the number of bins.

To get more accurate results, several formal tests of the Benford's law have been proposed in the literature (see, among others, [11–13]). However, it is a common experience that Benford datasets have been characterized by producing very small *p-values* of conformity tests even if the Benfordness null hypothesis is widely recognized to hold. A possible reason is that huge sample sizes make tests very powerful. Therefore, even small discrepancies from the Benford's law, affect the test significance. To avoid this disappointing feature, a strategy proposed by Barabesi et al. [14] consists of bootstrapping a number of smaller-sized samples, with the advice that a relatively high proportion of rejections cast doubt on the Benford nature of data. Additionally, small *p-values* could be commonly explained by the presence of some noise. This could be due to measurement errors. They arise, for example, when measurement devices are deteriorated or mis-calibrated or when, for some reason, it is difficult to measure the variables of interest.

In this paper, we pursue the density estimation task when Benford data are affected by measurement errors. In particular, we employ the deconvolution estimator, involving a standard kernel and a collection of higher order kernels. Then, after deconvolving the contaminated version of Benford data, we test a set of samples drawn from the resulting estimate to validate Benfordness. Ideally, we would like to

* Corresponding author.
E-mail address: chiara.passamonti@unich.it (C. Passamonti).

https://doi.org/10.1016/j.seps.2024.102008
Received 29 March 2024; Received in revised form 12 June 2024; Accepted 26 June 2024
Available online 29 June 2024
0038-0121/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

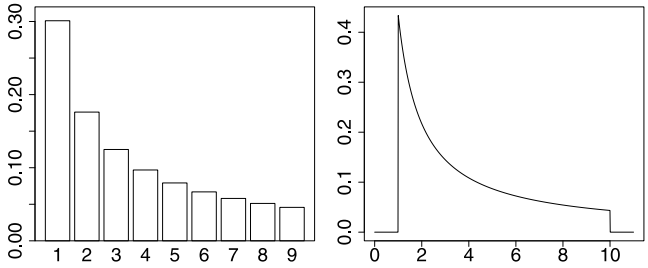


Fig. 1. The frequency distribution of the first significant digit (left). The probability density function of a Benford random variable ranging from 1 to 10 (right).

improve the discriminant power of tests, obtaining reasonably big *p-values* when data are Benford and small ones in the opposite situation.

In Section 2 we recall the main features of the Benford's law, and in Section 3 we discuss the deconvolution estimator. Some simulation results are presented in Section 4. Finally, in Section 5 we show a real data case study about the financial statement numbers of the Sino-Forest Corporation.

2. Benford's law

The frequency distribution of the first significant digit (f.s.d.) is well described by the following model, proposed by Newcomb [1]

$$P(f.s.d. = i) = \log_{10}\left(\frac{i+1}{i}\right) \tag{1}$$

where $i = 1, \dots, 9$. The frequency distribution of the f.s.d. is shown in the left panel of Fig. 1. An intuitive explanation of this law, based on arguments from circular statistics, follows. Assume that W is an absolutely continuous random variable. Consider the random variable $D = cW$, where c is a real number. A corollary of the Riemann Lebesgue theorem implies that the distribution of the wrapped variable $\lim_{c\to\infty} D(mod\,1)$ tends to be $U[0,1)$.

This result is an old one, as Poincaré [15] observed that the stopping position of a needle, which is free to rotate about the center of a disc, is uniformly distributed over different points on the circumference, if the total distance D covered by the rotations is described by a random variable having a large spread. Now, the random variable $Y = \log_{10} D$ can be represented on a circle of unitary circumference where each turn represents a unitary logarithm increment, and the stopping position $\theta = Y(mod\,1)$ is an angle that tends to be uniformly distributed because Y has a large spread.

Thanks to the uniformity of the stopping point θ, it is possible to derive the formula of the first significant digit distribution:

$$\begin{aligned} P(f.s.d. = i) &= P\left(i \times 10^k \le D < (i+1) \times 10^k\right) \\ &= P\left(\log_{10} i + k \le Y < \log_{10}(i+1) + k\right) \\ &= P\left(\log_{10} i \le \theta < \log_{10}(i+1)\right) \\ &\approx \log_{10}(i+1) - \log_{10} i \\ &= \log_{10}\left(\frac{i+1}{i}\right) \end{aligned} \tag{2}$$

where i ranges from 1 to 9 while k is an integer number.

It is possible to extend the above law to digits beyond the first. In particular, the probability that i, for $i = 0, 1, \dots, 9$, is encountered as the nth ($n > 1$) digit is

$$P(n^{th}s.d. = i) \sum_{k=10^{n-2}}^{10^{n-1}-1} \log_{10}\left(\frac{10k+i+1}{10k+i}\right).$$

In conclusion, a sequence of real numbers is Benford if and only if the decimal part of 10-based logarithm of its absolute value, the so-called *mantissa*, is uniformly distributed modulo one (see [16]).

Let U be a random variable uniformly distributed on $[0, 1)$. Then, the probability density function of the Benford random variable $X = 10^U$ is

$$f_X(x) = \frac{1}{x}\log_{10} e \qquad \text{for} \qquad x \in [1, 10). \tag{3}$$

(see [16] for more details, and the right panel of Fig. 1 for an illustration).

An important property related to the Benford's law is the *scale invariance* one. Pinkham [17] proved that if a collection of data follows the Benford's law, and we apply to all numbers the same positive constant, then, its re-scaled version is still conform to the Benford's law.

An interesting point, regarding the generation of Benford data, can be seen in the algorithm discussed in the following. Standard methods are not directly employable due to the fact that "Benford distribution" does not indicate any kind of parametric family. While, if Benford data are not considered in the original space, but in the mantissa one, they come from a continuous uniform density in [0,1).

1. Consider $M \ge 1$ densities whose mixture is continuous Uniform in [0,1) (for example consider an equal mixture between two Beta densities and a Uniform one: $\frac{1}{3}Be(1,2)+\frac{1}{3}U(0,1)+\frac{1}{3}Be(2,1)$). Draw from the ith density $n_i = n \times \omega_i$ observations, where ω_i is the weight of each component of the mixture, and $\sum_{i=1}^{M} n_i = n$.
2. Add the same integer number to all of the elements of the Mth sub-sample. In order to vary the magnitude of data we need to choose M distinct integers. Indeed, the magnitude of data is related to the *characteristic* of a number, i.e. the integer part of its 10-based logarithm. For example, numbers belonging to the intervals $[1, 10)$, $[10, 100)$, $[100, 1000)$ and $[1000, 10\,000)$ have, respectively, characteristics equal to 0, 1, 2 and 3.
3. Let T be the generic obtained observation, then the corresponding Benford datum is 10^T.

A typical, Benford dataset takes values in a compact range, and, consequently, kernel-based density estimators produce overflows outside of the support limits. Also, the density estimation task in the boundary region is further complicated by the fact that Benford data exhibit a mode at the left boundary, or very close to it, as shown in the right panel of Fig. 1.

In the following, we discuss the standard errors-in-variables problem and link this topic to Benford data.

3. Deconvolution kernel density estimation

Given a random sample $X_1, \dots, X_n$ from an unknown density f_X, the kernel estimator of f_X at x is defined as

$$\hat{f}_X(x; h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \tag{4}$$

where K is a kernel, i.e. a unimodal, symmetric function, that integrates to 1, and h is a non-negative smoothing parameter, called bandwidth.

Now, let $Y_1, \dots, Y_n$ be a random sample from an unknown density f_Y. Assume that

$$Y_i = X_i + \varepsilon_i, \qquad i = 1, \dots, n \tag{5}$$

where ε_i are independent copies of the random measurement error, independent of X, with a known probability density function, say f_ε, symmetric around zero. When the aim is to estimate the density f_X of unobservable random variable X, we are in the so-called errors-in-variables context. After noting that, according to model (5), f_Y is the convolution between f_X and f_ε, to recover f_X [18,19] proposed the deconvolution kernel density estimator.

Denote as ϕ_Z the characteristic function of a generic random variable Z. Given a random sample $Z_1, \dots, Z_n$, let $\hat{\phi}_Z(t) = n^{-1}\sum_{j=1}^{n} exp(itZ_j)$

be the empirical characteristic function of Z. Now, since $\phi_Y = \phi_X \phi_\varepsilon$ the deconvolution kernel estimator is

$$\tilde{f}_X(x;h) = \frac{1}{2\pi}\int exp(-itx)\frac{\hat{\phi}_Y(t)\phi_K(ht)}{\phi_\varepsilon(t)}dt = \frac{1}{nh}\sum_{i=1}^{n} L\left(\frac{Y_i - x}{h}\right), \tag{6}$$

where

$$L(x) = \frac{1}{2\pi}\int exp(-itx)\frac{\phi_K(t)}{\phi_\varepsilon(t/h)}dt \tag{7}$$

is the deconvolution kernel. The above estimator is well defined if the following conditions hold:

- (i) $\phi_\varepsilon(t) \neq 0$ for all $t \in R$
- (ii) $\int |\phi_X(t)|dt < \infty$
- (iii) $\sup_{t\in R}|\phi_K(t)/\phi_\varepsilon(t/h)| < \infty$ and $\int |\phi_K(t)/\phi_\varepsilon(t/h)|dt < \infty$.

In the context of kernel density estimation, it is commonly used a weight which is a unimodal probability density function, symmetric around zero. Alternatively, it is possible to employ some functions belonging to the class of higher order kernels. It is well known that, when K is not restricted to be a density function, and so a zero second moment of K is allowed, it is possible to get estimators with a reduced bias of magnitude $O(h^4)$. This leads to a faster convergence rate with respect to the classical kernel density estimator.

A collection of higher order kernels is presented in [20]. Starting from a symmetric, second order kernel K with $s_j = \int u^j K(u)du$, $j \geq 0$, the most used fourth order kernels are:

(a) the Lejeune and Sarda (LS) kernel

$$K^{LS}(u) := \frac{(s_4 - u^2 s_2)K(u)}{(s_4 - s_2^2)},$$

(b) the twicing (TW) kernel

$$K^{TW}(u) := 2K(u) - (K * K)(u), \tag{8}$$

(c) the Hall and Marron (HM) kernel

$$K^{HM}(u) := \frac{1}{2}\left(3K(u) + uK'(u)\right), \tag{9}$$

(d) the Jones and Foster (JF) kernel

$$K^{JF}(u) := K(u) - \frac{1}{2}s_2 K''(u).$$

which have been, respectively, obtained by Lejeune and Sarda [21], Stuetzle and Mittal [22], Hall and Marron [23] and Jones and Foster [24]. In the particular case where K is a standard normal density function K^{LS}, K^{HM} and K^{JF} are the same.

We employ the above higher order kernels in the deconvolution estimator (6), and discuss the advantages obtained in terms of asymptotic properties. For $s \in \{LS, TW, HM, JF\}$, let ϕ_{K^s} be the characteristic function of the kernel K^s. The higher order version of the deconvolution kernel L is obtained by using the ratio $\phi_{K^s}/\phi_\varepsilon$ in Eq. (7). Obviously, the use of all these kernels produces a $O(h^4)$ bias of the resulting kernel estimator.

Let $Y_1, \ldots, Y_n$ be a random sample generated by the model (5). When K is a kernel of order α with $s_0 = 1$, $s_j = 0$ for $1 < j < \alpha$ and $s_\alpha = c$ where $c > 0$ is some finite constant, then the following asymptotic bias and variance hold

$$\mathsf{E}[\tilde{f}_X(x,h)] - f_X(x) = \frac{h^\alpha s_\alpha}{\alpha!}f_X^\alpha(x) + o(h^\alpha), \tag{10}$$

$$\mathsf{VAR}[\tilde{f}_X(x,h)] = \frac{f_X(x)}{2\pi nh}\frac{|\phi_K(t)|^2}{|\phi_\varepsilon(t/h)|^2} + O(nh^{-1}) \tag{11}$$

Notice that, the bias of $\tilde{f}_X(\cdot,h)$ is the same as the bias of the error-free case estimator (4). This is not surprising because it can be shown that kernel (7) can be also obtained by the *unbiased score approach* (see [25]). While the spread of $\tilde{f}_X(\cdot,h)$ is also determined by the smoothness of f_ε which is said to be *supersmooth* if the rate of decay of $|\phi_\varepsilon(t)|$, as $|t| \to \infty$, is *exponential* (this is the case, for example, of the Normal and Cauchy densities), and it is *ordinary smooth* if $|\phi_\varepsilon(t)|$ tends to zero at a *polynomial* rate (this is the case, for example, of the Laplace and Gamma densities).

4. Simulations

In our experiment, we examine Benford data affected by measurement errors. We consider a very basic simulative scenario, where raw data range within a single magnitude order ($[1, 10)$), expressed as a power of ten. We extracted 500 samples of n observations from the Benford random variable $X = 10^U$ with U being the continuous Uniform density in $[0, 1]$, and $Y = X + \varepsilon$, where the measurement error is both Gaussian (super-smooth) or Laplace (ordinary-smooth), i.e. $\varepsilon \sim N(0, \sigma_\varepsilon)$ or $\varepsilon \sim L(0, \sigma_\varepsilon)$. For each sample we estimate f_X using estimator (6) where the deconvolution weight (7) employs, as the kernel function K, the standard Normal density or higher order kernels (8) and (9), respectively denoted as $\mathrm{DEC}_{N(0,1)}$, DEC_{TW} and DEC_{HM}. Then, we compare the results with the naive kernel density estimate (KDE). Our performance index is the averaged integrated squared error (AISE)

$$AISE(E) = \frac{1}{n}\sum_{i=1}^{500}\int (e_i(u) - f_X(u))^2 du$$

where E is either the naive or deconvolution estimators, $e_i(u)$ the corresponding density estimate at u using ith sample, and f_X is the target Benford probability density function reported in formula (3).

In order to isolate specific merits of the estimators, in this study we use for all the methods the best possible smoothing degree obtained through a grid search. We notice that, when a Gaussian error is assumed, for the deconvolution estimator the bandwidth of the kernel is constrained to be greater than the σ_ε.

Results are reported in Table 1. We observe the superiority of our estimators on the naive one. Also, the method appears to have a clear asymptotic nature, meaning that the relative performance improves with n.

A second way to evaluate performances is to compare *p-values* of a Benfordness test, in our case the Rayleigh one, conducted on the mantissas of $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$ and a sample of size n drawn from the deconvolution estimate. In Table 2 we report averaged *p-values* calculated using the samples of previous experiment. Clearly, the effect of measurement error is a strong decrease of *p-values*, while samples from the estimate exhibit an evident recover towards what was obtained from uncorrupted samples. Interestingly, when the kernel (9) is employed, the Rayleigh test exhibits averaged *p-values* very high even in correspondence of $n = 10\,000$. From a more general perspective, these simulative results say that, in presence of measurement errors, *p-values* decrease as n increase. This is in accordance with the very small *p-values* observed in the literature in the presence of big samples of Benford data with noise.

Particularly, we observe that higher-order methods tend to be superior. This is due to the fact that they are able to address the extra bias resulting from the presence of extra boundaries, more effectively considering their higher order nature. Performances shine when the error is Laplace. This is theoretically motivated by the fact that a super-smooth error density makes the estimation task much more challenging.

5. Real case study

Consolidated practical experience suggests that the first digits of data taken from a big number of balance sheets follow a Benford distribution, therefore a common application of Benford's law involves fraud detection by comparing the frequencies of leading digits in observed data with the theoretical ones. Nigrini [26] showed that Internal Revenue Service (IRS) tax data closely fit Benford's law, while

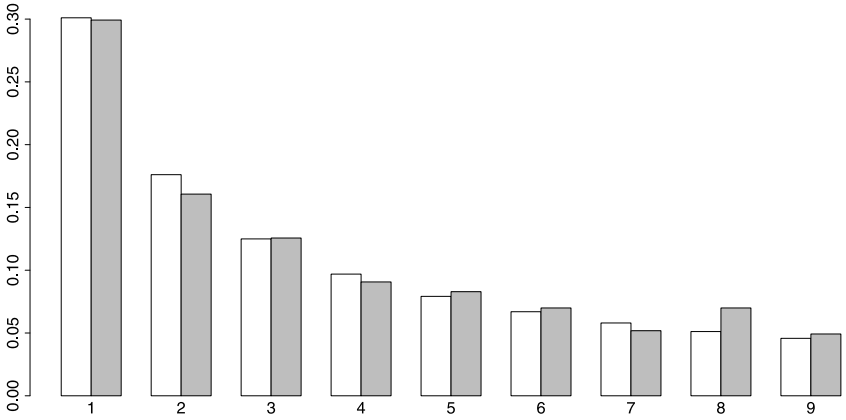


Fig. 2. Benford's law (white) and first digits of the 772 reported numbers (gray).

Table 1
AISE × 1000 of KDE, $\mathrm{DEC}_{N(0,1)}$, DEC_{TW} and DEC_{HM} over 500 samples of size n drawn from a Benford distribution contaminated with measurement error $\varepsilon \sim L(0, 0.4)$ ($\varepsilon \sim L(0, 0.5)$) (top panel) or $\varepsilon \sim N(0, 0.3)$ ($\varepsilon \sim N(0, 0.4)$) (bottom panel).

n	KDE	$\mathrm{DEC}_{N(0,1)}$	DEC_{TW}	DEC_{HM}
500	17.60 (20.04)	16.50 (18.26)	16.18 (17.80)	16.18 (17.80)
1000	15.98 (18.44)	14.16 (15.77)	13.72 (15.23)	13.70 (15.21)
5000	13.74 (16.34)	10.10 (11.47)	9.60 (10.88)	9.62 (10.90)
10 000	13.30 (15.94)	9.00 (10.23)	8.65 (9.70)	8.45 (9.60)
500	17.12 (20.33)	17.15 (20.35)	16.11 (17.83)	15.98 (18.47)
1000	15.54 (18.86)	15.58 (18.91)	14.37 (15.79)	13.97 (16.54)
5000	13.60 (17.14)	13.83 (17.41)	12.63 (15.35)	11.19 (13.84)
10 000	13.22 (16.87)	13.58 (17.23)	13.37 (14.90)	10.82 (13.50)

Table 2
Average of Rayleigh test *p-values* over 500 samples of size n drawn from Benford distribution (X), corrupted samples (Y) and samples drawn from the density estimates obtained using $\mathrm{DEC}_{N(0,1)}$, DEC_{TW} and DEC_{HM} when data are contaminated with measurement error $\varepsilon \sim L(0, 0.4)$ ($\varepsilon \sim L(0, 0.5)$) (top panel) or $\varepsilon \sim N(0, 0.3)$ ($\varepsilon \sim N(0, 0.4)$) (bottom panel).

n	X	Y	$\mathrm{DEC}_{N(0,1)}$	DEC_{TW}	DEC_{HM}
500	0.493	0.248 (0.157)	0.251 (0.214)	0.308 (0.275)	0.284 (0.275)
1000	0.469	0.124 (0.043)	0.234 (0.185)	0.283 (0.255)	0.261 (0.240)
5000	0.492	7e−04 (5e−06)	0.203 (0.141)	0.269 (0.234)	0.256 (0.211)
10 000	0.471	6e−05 (7e−09)	0.150 (0.087)	0.174 (0.162)	0.237 (0.197)
500	0.493	0.316 (0.179)	0.210 (0.139)	0.309 (0.224)	0.301 (0.250)
1000	0.469	0.220 (0.073)	0.174 (0.079)	0.270 (0.157)	0.277 (0.218)
5000	0.492	0.007 (2e−04)	0.008 (2e−05)	0.101 (0.019)	0.191 (0.064)
10 000	0.471	4e−04 (3e−09)	5e−04 (6e−08)	0.009 (0.002)	0.121 (0.017)

fraudulent data often do not. Carslaw [27] examined the second digit occurrence frequency of 220 New Zealand companies' balance sheets, and he noticed abnormalities regarding an excessive presence of the zero digit, and an occurrence of the nine digit lower than expected. This highlighted a manipulation aimed to round up companies' profits. Similarly, Thomas [28] examined U.S. COMPUSTAT firms to determine whether reported earnings followed similar patterns. He noticed that data revealed small deviations from expected frequencies, and reporting losses exhibited fewer zeros and more nines digits.

In the following, we discuss the Sino-Forest case, which provides a real-world example of financial misstatement and audit failure (see [29]). Sino-Forest Corporation was engaged primarily in the purchase and sale of standing timber in the People's Republic of China (PRC). As stated in the Sino-Forest 2011 annual report, its principal businesses include the ownership and management of plantation forests, the sale of standing timber and wood logs, and the complementary manufacturing of downstream engineered-wood products. The principal executive office was in Hong Kong and its securities were traded on the Toronto Stock Exchange until 2011. The company management established a complex system of subsidiaries and related entities to manage the buying and selling of standing timber across different regions of the People's Republic of China. However, Sino-Forest's employees created false documents to misrepresent these transactions, leading to inaccuracies in the company's financial reports. In June 2011, Carson Block, a private analyst and founder of Muddy Waters Research, exposed the fraud, causing the company's stock price to plummet. As a result, the company was delisted from the Toronto Stock Exchange and subsequently dissolved.

In the 2010 annual report, there were 772 financial statement numbers. These data are available in *benford.analysis* R package and they are discussed in [3]. The absolute value of negative numbers has been considered. The first digits, whose frequencies are depicted in Fig. 2, show an acceptable conformity to Benford's law, although the excess of the digit 8, which is a lucky number for the Chinese culture.

Since several irregularities have been identified in these statements (see, for example, [30]), we have some basis to model such observations as contaminated by an error source. Therefore we assume for these data model (5). Surely, in order to make the model compatible with an additive error generated by a single distribution, we need to restrict our attention to numbers ranging from 1000 to 10000, so considering 158 data within only one magnitude order (see Fig. 3).

Now, according to the result of the Rayleigh test performed on these data, it appears doubtful that a Benford hypothesis holds (*p-value* = 0.02). Therefore, since we are allowed to deconvolve these data, we use estimator (6) where the deconvolution weight (7) employs, as the kernel function K, the standard Normal density or higher order kernels (8) and (9), respectively denoted as $\mathrm{DEC}_{N(0,1)}$, DEC_{TW} and DEC_{HM}. The bandwidth has been selected by using the method proposed by Sheather and Jones [31]. As the distribution error, we use the Laplace one and, in order to make our conclusion somehow robust, we will try two different spread degrees.

In the first step, we compare the results with the naive kernel density estimate obtained using (4) in terms of integrated squared errors. From the results, shown in Table 3, we notice that the discrepancy between such estimates and the Benford model is higher for the standard method.

Additionally, in order to judge if the deconvolution task has been successfully accomplished, we draw a big number of samples of the same size as the original ones from the deconvolved estimates, and check if the averaged *p-values* are significantly bigger than 0.02. From results, reported in Table 4, we can see that the Rayleigh test on deconvoluted data provides averaged *p-values* significantly bigger in correspondence of both error spread levels, so conformity to Benford's law is now much clearer. Interestingly, in a scenario where the bias is abundant due to the presence of support boundaries at 1 and 10, higher order methods perform, in turn, much better than standard deconvolution due to their attitude to produce estimates with low bias.

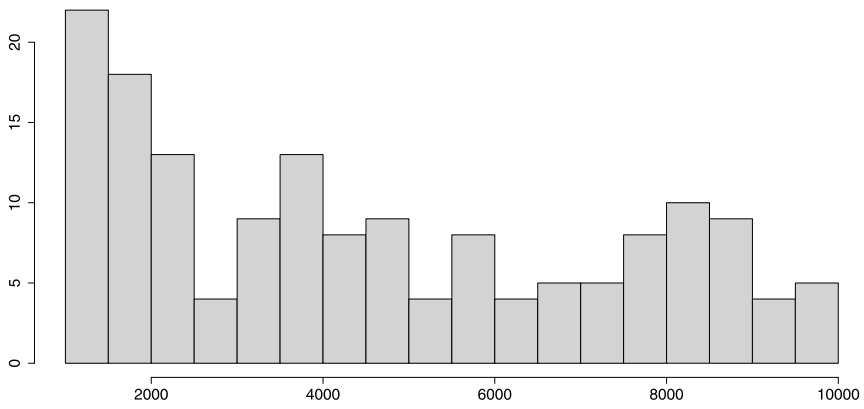


Fig. 3. Sino-Forest financial statements data ranging from 1000 to 10000.

Table 3
ISE × 1000 of density estimators on statement numbers assuming $\varepsilon \sim L(0, \sigma_\varepsilon)$.

σ_ε	KDE	$\text{DEC}_{N(0,1)}$	DEC_{TW}	DEC_{HM}
0.3	29.53	27.35	25.65	25.33
0.4	29.53	26.27	25.93	26.57

Table 4
p-values of the Rayleigh test on observed data (X) and averaged *p-values* over 500 samples of 150 observations drawn from deconvoluted estimates assuming $\varepsilon \sim L(0, \sigma_\varepsilon)$.

σ_ε	X	$\text{DEC}_{N(0,1)}$	DEC_{TW}	DEC_{HM}
0.3	0.022	0.051	0.065	0.072
0.4	0.022	0.047	0.074	0.074

6. Conclusions

A proposal to address challenges encountered in Benfordness statistical testing is presented, for the case where traditional methods may yield unexpected *p-values* due to potential contamination by noise. To mitigate these issues, we pursued a density estimation task taking into account the presence of measurement errors, using both classical and higher-order deconvolution kernels. Our approach, also in its lower-bias version, for tackling challenges in Benfordness statistical testing has filled a gap in the literature. A simulation study demonstrating the efficacy and relevance of the methodology is reported, and a real-data application relying on the Sino-Forest case is presented. To validate the efficacy of our procedure, additional datasets need to be analyzed. This task seems not to be so difficult because measurement errors commonly occur in real life.

CRediT authorship contribution statement

Marco Di Marzio: Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Stefania Fensore:** Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Chiara Passamonti:** Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data are available in the Benford.analysis package.

References

[1] Newcomb S. Note on the frequency of use of the different digits in natural numbers. Am J Math 1881;4(1):39–40.
[2] Benford F. The law of anomalous numbers. Proc Am Philos Soc 1938;551–72.
[3] Nigrini MJ. Benford's law: Applications for forensic accounting, auditing, and fraud detection, vol. 586, John Wiley & Sons; 2012.
[4] Deckert J, Myagkov M, Ordeshook PC. Benford's Law and the detection of election fraud. Political Anal 2011;19(3):245–68.
[5] Lacasa L, Fernández-Gracia J. Election forensics: Quantitative methods for electoral fraud detection. Forensic Scince Int 2019;294:19–22.
[6] Demir B, Javorcik B. Trade policy changes, tax evasion and Benford's law. J Dev Econ 2020;144:102456.
[7] Cerioli A, Barabesi L, Cerasa A, Menegatti M, Perrotta D. Newcomb–Benford law and the detection of frauds in international trade. Proc Natl Acad Sci 2019;116(1):106–15.
[8] Kolias P. Applying Benford's law to COVID-19 data: the case of the European Union. J Public Health 2022;44(2):221–6.
[9] Silva L, Figueiredo Filho D. Using Benford's law to assess the quality of COVID-19 register data in Brazil. J Public Health 2021;43(1):107–10.
[10] de Jong J, de Bruijne J, De Ridder J. Benford's law in the Gaia universe. Astron Astrophys 2020;642:A205.
[11] Kossovsky AE. Benford's law: theory, the general law of relative quantities, and forensic fraud detection applications, vol. 3, World Scientific; 2014.
[12] Barabesi L, Cerasa A, Cerioli A, Perrotta D. On characterizations and tests of Benford's law. J Amer Statist Assoc 2022;117(540):1887–903.
[13] Cerqueti R, Lupi C. Severe testing of Benford's law. Test 2023;32(2):677–94.
[14] Barabesi L, Cerioli A, Di Marzio M. Statistical models and the Benford hypothesis: a unified framework. TEST 2023;32(4):1479–507.
[15] Poincaré H. Chance. Monist 1912;31–52.
[16] Berger A, Hill TP. An introduction to Benford's law. Princeton University Press; 2015.
[17] Pinkham RS. On the distribution of first significant digits. Ann Math Stat 1961;32(4):1223–30.
[18] Carroll RJ, Hall P. Optimal rates of convergence for deconvolving a density. J Amer Statist Assoc 1988;83(404):1184–6.
[19] Stefanski LA, Carroll RJ. Deconvolving kernel density estimators. Statistics 1990;21(2):169–84.
[20] Jones MC, Signorini D. A comparison of higher-order bias kernel density estimators. J Amer Statist Assoc 1997;92(439):1063–73.
[21] Lejeune M, Sarda P. Smooth estimators of distribution and density functions. Comput Statist Data Anal 1992;14(4):457–71.
[22] Stuetzle W, Mittal Y. Some comments on the asymptotic behavior of robust smoothers. In: Smoothing Techniques for Curve Estimation: Proceedings of a Workshop Held in Heidelberg, April 2–4, 1979. Springer; 2006, p. 191–5.
[23] Hall P, Marron J. Variable window width kernel estimates of probability densities. Probab Theory Related Fields 1988;80(1):37–49.
[24] Jones M, Foster P. Generalized jackknifing and higher order kernels. J Nonparametr Stat 1993;3(1):81–94.
[25] Delaigle A. Nonparametric kernel methods with errors-in-variables: constructing estimators, computing them, and avoiding common mistakes. Aust N Z J Stat 2014;56(2):105–24.
[26] Nigrini MJ. The detection of income tax evasion through an analysis of digital distributions. University of Cincinnati; 1993.
[27] Carslaw CA. Anomalies in income numbers: Evidence of goal oriented behavior. Account Rev 1988;321–7.
[28] Thomas JK. Unusual patterns in reported earnings. Account Rev 1989;773–87.
[29] Wright GB, Cullinan C. Sino-forest corporation: the case of the standing timber. Glob Perspect Account Educ 2017;14:10–22.

[30] Mumic N, Filzmoser P. A multivariate test for detecting fraud based on Benford's law, with application to music streaming data. Stat Methods Appl 2021;30(3):819–40.
[31] Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. J R Stat Soc Ser B 1991;53:683—690.

Marco Di Marzio is full professor of Statistics at the University of Chieti-Pescara. He is a statistician with recognized contributions to statistical methods and applications in a wide range of domains. His main interests are in nonparametric density estimation, regression and classification, Machine Learning, Directional Statistics.

Stefania Fensore is research fellow of Statistics at the University of Chieti-Pescara. Her research topics include kernel methods, directional data and interdisciplinary research with special focus on health-related topics.

Chiara Passamonti is Ph.D. student in Human Sciences at the University of Chieti-Pescara. Her main interests are related to Benford data and errors-in-variables problems.