



# Statistical models and the Benford hypothesis: a unified framework

Lucio Barabesi<sup>1</sup> · Andrea Cerioli<sup>2</sup>  · Marco Di Marzio<sup>3</sup>

Received: 12 July 2023 / Accepted: 2 August 2023  
© The Author(s) 2023

## Abstract

The Benford hypothesis is the statement that a random sample is made of realizations of an absolutely continuous random variable distributed according to Benford's law. Its potential interest spans over many domains such as detection of financial frauds, verification of electoral processes and investigation of scientific measurements. Our aim is to provide a principled framework for the statistical evaluation of this statement. First, we study the probabilistic structure of many classical univariate models when they are framed in the space of the significand and we measure the closeness of each model to the Benford hypothesis. We then obtain two asymptotically equivalent and powerful tests. We show that the proposed test statistics are invariant under scale transformation of the data, a crucial requirement when compliance to the Benford hypothesis is used to corroborate scientific theories. The empirical advantage of the proposed tests is shown through an extensive simulation study. Applications to astrophysical and hydrological data also motivate the methodology.

**Keywords** Benford's law · Digit distribution · Goodness-of-fit · Kolmogorov distance · Likelihood ratio test · Significand

**Mathematics Subject Classification** 62E10 · 62F03 · 62G07 · 65C05

---

✉ Andrea Cerioli  
andrea.cerioli@unipr.it

Lucio Barabesi  
lucio.barabesi@unisi.it

Marco Di Marzio  
marco.dimarzio@unich.it

<sup>1</sup> Department of Economics and Statistics, University of Siena, Siena, Italy

<sup>2</sup> Department of Economics and Management, University of Parma, Parma, Italy

<sup>3</sup> Department of Philosophical, Pedagogical and Economic-Quantitative Sciences, "G. D'Annunzio" University, Pescara, Italy

## 1 Introduction

Benford's law is a much studied probability distribution for significant digits, which has been challenging mathematicians and practitioners for decades (Berger and Hill 2011, 2015; Miller 2015a). It has also attracted the interest of statisticians mainly thanks to the limit theorem presented by Hill (1995), which motivates the adoption of the law as the digit-generating model in many real-world situations. We then call *Benford hypothesis* the statement, to be made more precise in the following sections, that a random sample of  $n$  observations is made of realizations of an absolutely continuous random variable  $X$  which is distributed according to Benford's law.

Statistical assessment of the Benford hypothesis is crucial for many purposes. For instance, when the Benford hypothesis is expected to hold for genuine observations, deviations from it can be taken as evidence of possible data manipulation. A selection of effective developments of this idea can be found in a growing number of fields, including the analysis of financial statements (Tam Cho and Gaines 2007; Nigrini 2012), electoral processes (Mebane 2010; Pericchi and Torres 2011; Fernández-Gracia and Lacasa 2018) and international trade (Barabesi et al. 2018; Cerioli et al. 2019; Lacasa 2019; Barabesi et al. 2021). The Benford hypothesis may also help to corroborate or disprove scientific theories. Shao and Ma (2010) theoretically prove that the first-digit pattern complies with Benford's law for famous physical models such as the Boltzmann–Gibbs, the Fermi–Dirac and the Bose–Einstein statistics. Nigrini and Miller (2007) check the conformity of streamflow data arising in hydrology and conclude that data related to water bodies should adhere to the Benford hypothesis, except in cases where the data quality is poor, or the sampling mechanism is flawed, or the hydrological process follows a power law with a large exponent. In astrophysics, a wealth of research has been recently conducted in order to address the question whether the distribution of the digits of physical quantities of celestial bodies is in good agreement with Benford's law (Alexopoulos and Leontsinis 2014; de Jong et al. 2020; Melita and Miraglia 2021). In the case of distances from the Earth to galaxies, Hill and Fox (2016) theoretically explain—using astrophysical arguments and the mathematical properties of Benford's law—why these distances should follow the Benford hypothesis and conclude that empirical observation of such an agreement may be viewed as a new independent evidence of the validity of Hubble's law. Accurate statistical assessment of the Benford hypothesis would then be required, but most of the applied astrophysical literature either rely on diagnostic checks of the data or just perform statistical evaluation of the looser hypothesis of first-digit compliance (see, e.g., Melita and Miraglia 2021).

Formal tests of the Benford hypothesis already exist and have been widely debated in the literature: see, e.g., Kossovsky (2015, Section 3) for a general account and Barabesi et al. (2022), Cerasa (2022) and Cerqueti and Lupi (2023) for a few recent proposals. Notwithstanding the difficulty of separating statistical significance from practical importance of an empirical observed deviation from Benford's law, which implies that formal tests might be judged too severe in rejecting the Benford hypothesis especially in large samples (Kossovsky 2015), in this work we stress that most existing tests suffer from two main drawbacks. The first shortcoming is that they are rather ad-hoc in nature, being derived on the basis of some specific properties of Benford's law.

**Table 1** Synthetic data set of  $n = 30$  observations

101	102	103	131	145	148	150	155	165	185
242	245	246	267	268	299	329	341	356	363
393	404	432	471	521	586	632	781	856	932

Each property may yield good statistical performance under certain alternatives, but poorer behavior in other circumstances; see Barabesi et al. (2022) for a detailed study and comparison of two such properties. As a consequence, a uniformly most powerful test does not exist. The second drawback of the available test statistics is their lack of scale invariance, except for the trivial case of multiplication by a power of 10. For example, consider the synthetic data set of  $n = 30$  observations displayed in Table 1. On these data the classical first-digit Pearson statistic, which is introduced in Sect. 4.1 and there denoted by  $\chi^2$ , equals 1.79 and yields an exact  $p$ -value of 0.998. If all the observations in the table are multiplied by the fixed scaling factor  $\sigma = 373$ , we instead obtain  $\chi^2 = 16.73$  and an exact  $p$ -value of 0.036. The same happens with another popular first-digit discrepancy measure, the mean absolute deviation also introduced in Sect. 4.1 and there denoted by  $M$ , for which we obtain  $M = 0.022$  (with an exact  $p$ -value of 0.984) on the data of Table 1, and  $M = 0.055$  (with an exact  $p$ -value of 0.147) after scaling by  $\sigma$ . While lack of invariance may be tolerable to some extent in certain domains, it has certainly to be avoided for the purpose of validating scientific theories and, more generally, in the analysis of scientific data.

The main goal of this paper is to provide a principled and unified framework for the assessment of the Benford hypothesis. Our aim is achieved by taking two alternative but connected routes that originate from the same source: the characteristic function of a suitable transform of the (univariate) random variable which is assumed to define the data generating process. One path leads to measure how close is the statistical model of interest from the Benford hypothesis, while the second track takes an inferential perspective and develops powerful and scale-invariant test statistics of this hypothesis.

Specifically, in the first part of the manuscript we develop a general relationship between the distribution of  $X$  and that of its significant  $S(X)$ , to be defined in Sect. 2. This relationship allows us to study the probabilistic structure of many standard univariate models for  $X$  when they are framed in the significant space. Although a number of scattered results exist on this topic (see, e.g., Engel and Leuenberger 2003; Nigrini and Miller 2007; Dümbsgen and Leuenberger 2008; Miller 2015b; Leemis 2015; Dümbsgen and Leuenberger 2015; Berger and Twelves 2018), most of the findings reported in Sect. 3 are new. Even more importantly, our approach is general, being based on the study of the characteristic function of a convenient transform of  $X$ . It is thus readily available for potential extension to other univariate models not considered here and for developments in more complex schemes, such as mixtures and contamination models. We establish the closeness of a particular model to the Benford hypothesis through a suitable Kolmogorov distance, and we suggest an easy strategy for regularization. The proposed metric is a natural choice in our framework and could be potentially extended to deal with outlier contamination in situations where robust validation and testing are of interest (Álvarez-Esteban et al. 2012; del Barrio et al. 2020). We argue

that providing a general relationship between the distribution of  $X$  and that of its significant  $S(X)$  can be fruitful for two orders of reasons. One is that it helps to translate much of the well-understood knowledge on  $X$  into the more slippery world of the significant space (see Berger and Hill 2021 for a list of surprisingly common mistakes in the field). The second advantage is that the computed discrepancy between  $X$  and the Benford hypothesis helps to sharpen the meaning of what might be called “almost Benford” behavior (Demidenko 2020, Section 2.1.6). Most real-world data generating processes do not exactly follow Benford’s law, although it is known that some of them may closely resemble the law for certain parameter configurations; see, e.g., Berger and Hill (2015, p. 55), Miller (2015b, Section 3.5.3) and Cerqueti and Maggi (2021). Therefore, our discrepancy measure allows us to represent deviations from the Benford hypothesis in terms of more comfortable and understandable variations in the parameters of  $X$ .

The relationship between the distributions of  $X$  and  $S(X)$  leads, in the second part of our work, to effective test statistics for assessing the Benford hypothesis when a random sample from  $X$  is available. Our approach takes advantage of a density estimation step based on a possibly unknown number of nonnegative trigonometric polynomials. Under this approach, we obtain, in Sect. 4, two test statistics of the Benford hypothesis which are shown to be asymptotically equivalent. One of them provides the so far missing likelihood ratio test of the Benford hypothesis, while the other statistic is as an extension of the well-known Rayleigh test which is effective against multimodal alternatives. We allow for a data-driven selection of the number of nonnegative trigonometric polynomials and we show that the proposed test statistics are invariant under scale transformation of the data.

In Sect. 5 the empirical advantage of the proposed tests of the Benford hypothesis is shown through an extensive simulation study encompassing many classical models for  $X$ . Two applications of scientific interest, one to astrophysical data and the other to hydrological data, conclude the paper in Sects. 6 and 7, respectively. These applications further motivate our developments and clearly point to the requirement of scale-invariant test statistics. All the technical proofs are deferred to the Appendix, available through the Supplementary Materials. There, we also provide a Supplement with additional simulation results, details about the simulation algorithm and links both to the code that we have used for our computations and to the data on which our tests are performed.

## 2 Preliminaries

We provide some basic definitions and properties of Benford’s law that are required by the developments of the following sections.

The (base-10) significant function  $S : \mathbb{R} \setminus \{0\} \rightarrow [1, 10[$  is defined as

$$S(x) = 10^{(\log_{10}|x|)},$$

where  $\langle a \rangle = a - [a]$ , and  $[a] = \max\{n \in \mathbb{Z} : n \leq a\}$ . Let  $X$  be an absolutely continuous random variable defined on the probability space  $(\Omega, \mathcal{F}, P)$ . According

to Berger and Hill (2015, p.30),  $X$  is a Benford random variable if the distribution function of  $S(X)$  is

$$F_{S(X)}(u) = \log_{10} u, \quad u \in [1, 10[. \tag{1}$$

We say that the distribution of  $X$  satisfies the *Benford hypothesis* when (1) holds.

In our perspective it is often convenient to refer to the (base-10) log-significand function

$$s(x) = \langle \log_{10}|x| \rangle,$$

with the assumption that  $s(0) := 0$ . If  $F_{s(X)}$  is the distribution function of  $s(X)$ , the Benford hypothesis is then equivalently stated as

$$F_{s(X)}(u) = u, \quad u \in [0, 1[. \tag{2}$$

We are interested in providing a bridge between the usual statistical approach, where a model is placed on the distribution of  $X$ , and the Benford hypothesis. For this purpose, we acknowledge the contribution of Pinkham (1961), who provided a pioneering theoretical discussion of why and to what extent Benford’s law must hold. The key finding given by Pinkham (1961) consists in the following Theorem.

**Theorem 1** *Let  $X$  be an absolutely continuous random variable defined on the probability space  $(\Omega, \mathcal{F}, P)$ . Moreover, let  $\varphi_Y$  be the characteristic function of the random variable  $Y = 2\pi \log_{10}|X|$ , i.e.,*

$$\varphi_Y(t) = \mathbb{E} \left[ e^{2\pi i t \log_{10}|X|} \right], \quad t \in \mathbb{R}. \tag{3}$$

If  $\varphi_Y(t) = O(|t|^{-h})$  when  $|t| \rightarrow \infty$  for a given  $h > 0$ , we have

$$F_{s(X)}(u) = u + \frac{1}{2\pi} \sum_{k \in \mathbb{Z}, k \neq 0} \frac{1 - e^{-2\pi i k u}}{i k} \varphi_Y(k), \quad u \in [0, 1[. \tag{4}$$

**Proof** See Pinkham (1961). □

An alternative (and more manageable) expression for (4) is introduced in the following corollary.

**Corollary 1** *We have*

$$F_{s(X)}(u) = u + \frac{1}{\pi} \sum_{k \in \mathbb{N}} \frac{|\varphi_Y(k)|}{k} (\sin(\vartheta_k) - \sin(\vartheta_k - 2\pi k u)), \quad u \in [0, 1[, \tag{5}$$

where  $\vartheta_k = \arg(\varphi_Y(k))$ .

**Proof** See the Appendix in the Supplementary Materials. □

It is apparent from Corollary 1 that the coefficients  $(|\varphi_Y(k)|)_{k \geq 1}$  are central, since they determine the series representation (5). Moreover, the  $N$ th partial sum of (5) is denoted by

$$G_N(u) = u + \frac{1}{\pi} \sum_{k=1}^N \frac{|\varphi_Y(k)|}{k} (\sin(\vartheta_k) - \sin(\vartheta_k - 2\pi ku)), \quad u \in [0, 1[. \quad (6)$$

A series representation for the probability density function of  $s(X)$ , say  $f_{s(X)}$ , is also obtained in the following corollary, albeit this result is not adopted in the sequel of the paper.

**Corollary 2** *If  $\sum_{k \in \mathbb{N}} |\varphi_Y(k)| < \infty$ , we have*

$$f_{s(X)}(u) = 1 + 2 \sum_{k \in \mathbb{N}} |\varphi_Y(k)| \cos(\vartheta_k - 2\pi ku), \quad u \in [0, 1[.$$

**Proof** See the Appendix in the Supplementary Materials. □

For instance, the condition in Corollary 2 is satisfied when  $|\varphi_Y(k)| = O(k^{-h})$  with  $h > 1$  or when  $\varphi_Y$  is absolutely integrable, since  $\sum_{k \in \mathbb{N}} |\varphi_Y(k)| \leq \int_{-\infty}^{\infty} |\varphi_Y(t)| dt$ . Actually, the common parametric families of absolutely continuous circular densities (considered, e.g., by Mardia and Jupp 2000) satisfy such an assumption. Moreover, almost all the laws considered in this paper also satisfy the condition, the only exceptions being the Beta distribution with shape parameter  $\beta \leq 1$  and the generalized Benford distribution.

We measure the closeness of the distribution of  $X$  to the Benford hypothesis (2) through the Kolmogorov distance between the distribution function of the uniform law on  $[0, 1[$  and (6) for  $N \in \mathbb{N}$ . This distance is

$$\begin{aligned} \Delta_N &= \sup_{u \in [0, 1[} |G_N(u) - u| \\ &= \sup_{u \in [0, 1[} \left| \frac{1}{\pi} \sum_{k=1}^N \frac{|\varphi_Y(k)|}{k} (\sin(\vartheta_k) - \sin(\vartheta_k - 2\pi ku)) \right|. \end{aligned} \quad (7)$$

It is apparent that  $\Delta_N$  is completely determined by the vector  $(\varphi_Y(1), \dots, \varphi_Y(N))$ . Furthermore, the Kolmogorov distance (7) provides a further equivalent statement of the Benford hypothesis as  $\Delta_N = 0$  for each  $N \in \mathbb{N}$ .

Another recurrent theme of our work is the effect of scaling. Let  $X_\sigma = \sigma X$  denote the scaled version of the random variable  $X$  for  $\sigma \in \mathbb{R}^+$ . A basic and easily seen property of the significand is that  $F_{s(X_\sigma)}$  has a periodic behavior with respect to the scale parameter  $\sigma$ , i.e.,

$$F_{s(X_\sigma)} = F_{s(X_{10^n \sigma})}, \quad n \in \mathbb{N}. \quad (8)$$

We expect this feature to be retrieved from the behavior of the Kolmogorov distance (7). However, as already anticipated in Sect. 1, it does not automatically translate into a sensible invariance property when testing conformance to (1) and (2).

We finally introduce the symbols  $\Delta_\infty = \lim_N \Delta_N$  and  $C = \log_{10} e$ , to be used in the developments that follow.

### 3 Statistical models under the Benford lens

In this section we build on (3) and (5) to assess the structure of the coefficients  $(|\varphi_Y(k)|)_{k \geq 1}$  for some well-known models for  $X$ , which are representative of a wide range of probabilistic behavior. It is clear from (5) that these coefficients have a key role in assessing the discrepancy of the distribution of a given random variable from the distribution of a Benford random variable. For well-behaved statistical models, the first few coefficients should reasonably “characterize” the underlying distribution in the Benford’s framework, in such a way that  $G_N$  provides a suitable approximation of  $F_{S(X)}$  even for a small  $N$ . Therefore, an appropriate analysis of these quantities is required.

Our aim is to recast the structure of some basic models for  $X$  in terms of their agreement with the Benford hypothesis. We argue that these relationships can help to better link the usual approach where the focus is on  $X$  to the alternative statistical view where the reference space is that of  $S(X)$  (or, equivalently, that of  $s(X)$ ). The reward of such a link is twofold. On the one hand, we can better understand the instances where the data generating process may be expected to be close to (or far from) the Benford behavior by relating this behavior to statistical models that are well understood after many decades (and even centuries) of research and empirical applications. On the other hand, if we trust a statistical model for  $X$  to be adequate for our problem of interest, we should find the corresponding information also in the significand space. When it is not the case, suspicion on the data should then arise. Such a dual role of the statistician, looking both at the observed value of  $X$  and at its digits, is also envisaged by Kossovsky (2015, p. 3–4) as a herald of fruitful statistical investigations. We believe that the results provided in this section could be helpful in this direction as well.

The ratio behind our selection of distributions stems from the well-known popularity of these probability laws. The choice was also carried out in order to cover a wide range of values with respect to the skewness and kurtosis indexes. Moreover, despite their central importance, we are not aware of existing results on the structure of the coefficients  $(|\varphi_Y(k)|)_{k \geq 1}$  for the normal and positive stable distributions, as well as for mixture of laws.

#### 3.1 Normal law

We start our journey under the Benford umbrella from the ubiquitous normal law.

**Proposition 1** *If the random variable  $X$  follows the normal law  $N(\mu, \sigma^2)$ , then*

$$\varphi_Y(t) = \frac{1}{\sqrt{\pi}} (2\sigma^2)^{\pi i C t} \Gamma\left(\frac{1}{2} + \pi i C t\right) M\left(-\pi i C t, \frac{1}{2}, -\frac{\mu^2}{2\sigma^2}\right), \quad t \in \mathbb{R},$$

while

$$|\varphi_Y(t)| = \frac{1}{\sqrt{\cosh(\pi^2 C t)}} \left| M\left(-\pi i C t, \frac{1}{2}, -\frac{\mu^2}{2\sigma^2}\right) \right|,$$

where  $M(a, b, z)$  is the Kummer confluent hypergeometric function and  $a, b, z \in \mathbb{C}$ .

**Proof** See the Appendix in the Supplementary Materials.  $\square$

Noteworthy, simple expressions can be derived from Proposition 1 when  $\mu = 0$ , since  $M(a, b, 0) = 1$  in that case. It is also apparent that the coefficients  $(|\varphi_Y(k)|)_{k \geq 1}$  are solely functions of the ratio  $\rho = |\mu|/\sigma$ . Numerical evaluation of these coefficients is provided in Table 2 for  $k \leq 5$  and selected values of  $\rho$ . We can see that the coefficients are increasing (and approach one) for a fixed  $k$  as  $\rho$  increases, while they are decreasing (and approach zero) for a fixed  $\rho$  as  $k$  increases. In addition, the Kolmogorov distance  $\Delta_N$  is reported in Table 3 for  $N \leq 5$  and some pairs  $(\mu, \sigma)$ . These distances are fluctuating for a fixed  $\sigma$  and for a varying  $\mu$ , while they tend to decrease as  $\sigma$  increases for a fixed  $\mu$ . They also have a periodic behavior with respect to the scale parameter  $\sigma$  for a fixed  $\rho$ , in agreement with (8). Finally, from the joint analysis of Tables 2 and 3, it can be seen that a good approximation to  $F_{s(X)}$  is obtained by truncating the series in (5) at the first term.

### 3.2 Gamma law

Another long-standing distribution with fruitful applications in many domains, including survival analysis, actuarial science and economics, is the Gamma law.

**Proposition 2** *If the random variable  $X$  follows the Gamma law  $G(\alpha)$ , where  $\alpha$  is the shape parameter, then*

$$\varphi_Y(t) = \frac{\Gamma(\alpha + 2\pi i C t)}{\Gamma(\alpha)}, \quad t \in \mathbb{R},$$

while

$$|\varphi_Y(t)| = \prod_{n=0}^{\infty} \frac{1}{\sqrt{1 + \frac{4\pi^2 C^2 t^2}{(\alpha+n)^2}}}.$$

**Proof** See the Appendix in the Supplementary Materials.  $\square$



**Table 2** Coefficients  $|\varphi_Y(k)|$  for  $k \leq 5$  and selected parameters of one-population laws

$k$	Normal law $N(\mu, \sigma^2)$ , with $\rho =  \mu /\sigma$					
	$\rho = 0$	$\rho = 1$	$\rho = 2$	$\rho = 3$	$\rho = 5$	
1	0.1658	0.2261	0.4445	0.6407	0.8506	
2	0.0195	0.0418	0.1251	0.2575	0.5470	
3	0.0023	0.0071	0.0295	0.0830	0.2859	
4	0.0003	0.0011	0.0063	0.0234	0.1281	
5	0.0000	0.0002	0.0013	0.0060	0.0511	
$k$	Gamma law $G(\alpha)$					
	$\alpha = 1.0$	$\alpha = 1.5$	$\alpha = 2.0$	$\alpha = 3.0$	$\alpha = 4.0$	
1	0.0570	0.1079	0.1655	0.2800	0.3785	
2	0.0011	0.0029	0.0061	0.0179	0.0371	
3	0.0000	0.0001	0.0002	0.0006	0.0019	
4	0.0000	0.0000	0.0000	0.0000	0.0001	
5	0.0000	0.0000	0.0000	0.0000	0.0000	
$k$	Beta law $Be(\alpha, \beta)$					
	$\beta = 1.0$			$\beta = 2.0$		
	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 1.5$	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 1.5$
1	0.1802	0.3441	0.4817	0.0868	0.2034	0.3254
2	0.0912	0.1802	0.2650	0.0242	0.0620	0.1104
3	0.0610	0.1213	0.1802	0.0110	0.0288	0.0526
4	0.0458	0.0912	0.1361	0.0062	0.0164	0.0304
5	0.0366	0.0731	0.1093	0.0040	0.0106	0.0197
$k$	Positive stable law $PS(\alpha)$					
	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.9$	
1	0.0000	0.0001	0.0195	0.1904	0.6547	
2	0.0000	0.0000	0.0003	0.0303	0.4066	
3	0.0000	0.0000	0.0000	0.0048	0.2526	
4	0.0000	0.0000	0.0000	0.0008	0.1569	
5	0.0000	0.0000	0.0000	0.0001	0.0974	

Some simplifications occur in special cases. As an example, since for  $b \in \mathbb{R}$  it holds

$$|\Gamma(1 + ib)| = \sqrt{\frac{\pi b}{\sinh(\pi b)}}, \tag{9}$$

**Table 3** Kolmogorov distance  $\Delta_N$  for the laws of Table 2 and  $N \leq 5$ 

$N$	Normal law $N(\mu, \sigma^2)$					
	$\sigma = 1$			$\sigma = 5$		
	$\mu = 0$	$\mu = 3$	$\mu = 5$	$\mu = 0$	$\mu = 3$	$\mu = 5$
1	0.0547	0.2324	0.5244	0.1035	0.1120	0.1266
2	0.0607	0.2651	0.5097	0.1035	0.1116	0.1220
3	0.0605	0.2629	0.4786	0.1038	0.1117	0.1222
4	0.0605	0.2655	0.4753	0.1038	0.1117	0.1221
5	0.0605	0.2656	0.4770	0.1038	0.1117	0.1221
$N$	Gamma law $G(\alpha)$					
	$\alpha = 1.0$	$\alpha = 1.5$	$\alpha = 2.0$	$\alpha = 3.0$	$\alpha = 4.0$	
1	0.0307	0.0683	0.1009	0.1085	0.1803	
2	0.0305	0.0682	0.1015	0.1135	0.1704	
3	0.0305	0.0682	0.1015	0.1135	0.1704	
4	0.0305	0.0682	0.1015	0.1135	0.1704	
5	0.0305	0.0682	0.1015	0.1135	0.1704	
$N$	Beta law $Be(\alpha, \beta)$					
	$\beta = 1.0$			$\beta = 2.0$		
	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 1.5$	$\alpha = 0.5$	$\alpha = 1.0$	$\alpha = 1.5$
1	0.1138	0.2123	0.2877	0.0451	0.1187	0.2017
2	0.1170	0.2259	0.3169	0.0437	0.1140	0.1966
3	0.1265	0.2362	0.3206	0.0438	0.1176	0.2041
4	0.1288	0.2458	0.3405	0.0442	0.1173	0.2027
5	0.1311	0.2458	0.3443	0.0441	0.1174	0.2043
$N$	Positive stable law $PS(\alpha)$					
	$\alpha = 0.1$	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.9$	
1	0.0000	0.0001	0.0100	0.1204	0.2747	
2	0.0000	0.0001	0.0101	0.1206	0.3770	
3	0.0000	0.0001	0.0101	0.1215	0.4165	
4	0.0000	0.0001	0.0101	0.1215	0.4244	
5	0.0000	0.0001	0.0101	0.1214	0.4218	

for the Exponential distribution ( $\alpha = 1$ ) we have

$$|\varphi_Y(t)| = \sqrt{\frac{2\pi^2 Ct}{\sinh(2\pi^2 Ct)'}}$$

on the basis of Olver et al. (2010, Formula 5.4.3). Similar closed-form expressions can be obtained for  $\alpha = n + \frac{1}{2}$  and  $\alpha = n + 1$ , with  $n \in \mathbb{N}$ , by considering the properties of the Gamma function.

Numerical computation of the coefficients  $(|\varphi_Y(k)|)_{k \geq 1}$  for  $k \leq 5$  is again provided in Table 2 for selected values of  $\alpha$ . For a fixed  $k$ , the coefficients are increasing (and approach one) as  $\alpha$  increases, while they are decreasing (and approach zero) as  $k$  increases. Table 3 displays the Kolmogorov distances  $\Delta_N$ , which consistently increase with  $\alpha$ . Finally, it is seen that also for the Gamma law a good approximation to  $F_s(X)$  can be achieved by truncating the series in (5) at the first term.

### 3.3 Beta law

The Beta law is a classical model for random variables that have a finite range.

**Proposition 3** *If the random variable  $X$  follows the Beta law  $Be(\alpha, \beta)$ , where  $\alpha$  and  $\beta$  are the shape parameters, then*

$$\varphi_Y(t) = \frac{\Gamma(\alpha + \beta)\Gamma(\alpha + 2\pi i C t)}{\Gamma(\alpha)\Gamma(\alpha + \beta + 2\pi i C t)}, \quad t \in \mathbb{R},$$

while

$$|\varphi_Y(t)| = \prod_{n=0}^{\infty} \sqrt{\frac{1 + \frac{4\pi^2 C^2 t^2}{(\alpha + \beta + n)^2}}{1 + \frac{4\pi^2 C^2 t^2}{(\alpha + n)^2}}}.$$

**Proof** See the Appendix in the Supplementary Materials. □

Similarly to the Gamma law, the expression of  $|\varphi_Y(t)|$  can be simplified in some special cases. For the uniform distribution it reduces to

$$|\varphi_Y(t)| = \sqrt{\frac{1}{1 + 4\pi^2 C^2 t^2}}.$$

If  $\alpha = \beta = \frac{1}{2}$ , i.e., when the arcsine distribution is considered, we obtain

$$|\varphi_Y(t)| = \sqrt{\frac{\tanh(2\pi^2 C t)}{2\pi^2 C t}},$$

since for  $b \in \mathbb{R}$  it holds

$$\left| \frac{\Gamma(\frac{1}{2} + ib)}{\Gamma(1 + ib)} \right| = \sqrt{\frac{\tanh(\pi b)}{b}}.$$

Moreover, closed-form expressions for  $|\varphi_Y(t)|$  can be derived when  $\alpha$  and  $\beta$  take the values  $n + \frac{1}{2}$  or  $n + 1$ , with  $n \in \mathbb{N}$ .

We give the coefficients  $(|\varphi_Y(k)|)_{k \geq 1}$  for  $k \leq 5$  in Table 2, and the Kolmogorov distances  $\Delta_N$  in Table 3, for selected pairs  $(\alpha, \beta)$ . For a fixed  $k$ , the coefficients are increasing (and approach one) as  $\alpha$  increases for a fixed  $\beta$ , they are decreasing (and approach zero) as  $\beta$  increases for a fixed  $\alpha$ , while they are decreasing (and approach zero) as  $k$  increases. Also for the Beta model, the final consideration is that a good approximation to  $F_{s(X)}$  is reached by truncating the series in (5) at the first term, or at most at the second one.

### 3.4 Positive stable law

We consider the positive stable law with Laplace transform  $L_X(t) = e^{-t^\alpha}$ , for  $\Re(t) \in \mathbb{R}^+$ , as a representative of the class of skewed and heavy-tailed laws.

**Proposition 4** *If the random variable  $X$  follows the positive stable law  $PS(\alpha)$ , where  $\alpha$  is the tail parameter, then*

$$\varphi_Y(t) = \frac{\Gamma\left(1 - \frac{2\pi i C t}{\alpha}\right)}{\Gamma(1 - 2\pi i C t)}, \quad t \in \mathbb{R},$$

while

$$|\varphi_Y(t)| = \frac{\sqrt{\sinh(2\pi^2 C t)}}{\alpha \sinh\left(\frac{2\pi^2 C t}{\alpha}\right)}.$$

**Proof** See the Appendix in the Supplementary Materials.  $\square$

The coefficients  $(|\varphi_Y(k)|)_{k \geq 1}$  for  $k \leq 5$  are displayed in Table 2 for selected values of  $\alpha$ . For a fixed  $k$ , they are increasing with  $\alpha$ , while they are decreasing (and approach zero) as  $k$  increases. Table 3 reports the values of the Kolmogorov distances  $\Delta_N$ , which steadily increase with  $\alpha$ . Finally, also in the present model a suitable approximation to  $F_{s(X)}$  is achieved by truncating the series in (5) at the first couple of terms.

### 3.5 Mixtures

A more complicated situation of interest arises when  $X$  is distributed as a mixture of  $L$  laws with normalized weights  $(w_1, \dots, w_L)$ , so that its probability density is  $f_X = \sum_{l=1}^L w_l f_{Z_l}$ , where  $f_{Z_l}$  is the density of  $Z_l$ . Mixture models have gained huge popularity, since they can be seen as the building blocks of both model-based clustering and outlier-contamination models (see, e.g., Farcomeni and Punzo 2020; Hennig 2022; Ingrassia et al. 2022). In the case of a mixture, if  $V_l = 2\pi \log_{10}|Z_l|$  and  $\varphi_{V_l}$  is the characteristic function of  $V_l$ , the characteristic function of  $Y = 2\pi \log_{10}|X|$  is given by

$$\varphi_Y(t) = \sum_{l=1}^L w_l \varphi_{V_l}(t). \quad (10)$$

**Table 4** As Table 2, but now for the normal mixtures (a)–(f)

$k$	(a)	(b)	(c)	(d)	(e)	(f)
1	0.0537	0.0476	0.0394	0.0779	0.0711	0.1985
2	0.1441	0.2820	0.1843	0.2125	0.2045	0.2667
3	0.0076	0.0148	0.0110	0.0358	0.0270	0.0722
4	0.0081	0.0284	0.0121	0.0137	0.0147	0.0334
5	0.0004	0.0016	0.0010	0.0038	0.0029	0.0101

It is difficult to envisage the effect that the components of the convex combination exert on the distribution of  $s(X)$ , as simple relationships between  $|\varphi_Y(k)|$  and  $|\varphi_{V_1}(k)|, \dots, |\varphi_{V_L}(k)|$  do not exist. Result (10) thus anticipates a general lack of predictability in the power results under a mixture model for  $X$ . Nevertheless, an amusing specialization of (10) occurs if we assume a location-scale family for the distribution of  $Z_l$  such that the ratio of the location parameter to the scale parameter is given by the fixed constant  $\rho = \frac{\mu}{\sigma}$ .

**Proposition 5** Let  $f_Z$  be the probability density of a standard random variable  $Z$ , and let  $(Z_1, \dots, Z_L)$  be absolutely continuous random variables with probability densities

$$f_{Z_l}(x) = \frac{1}{k_l \sigma} f_Z\left(\frac{x - k_l \mu}{k_l \sigma}\right), \quad l = 1, \dots, L,$$

where  $\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$  and  $k_l > 0$ . If  $X$  is a mixture with probability density  $f_X = \sum_{l=1}^L w_l f_{Z_l}$ , then

$$\varphi_Y(t) = \varphi_{Y_1}(t)\varphi_{Y_2}(t), \quad t \in \mathbb{R},$$

where  $Y_1 = 2\pi \log_{10}|X_1|$  and  $Y_2 = 2\pi \log_{10}|X_2|$ , while  $X_1$  and  $X_2$  are independent random variables such that  $X_1$  is discrete with  $P(X_1 = k_l) = w_l$ , for  $l = 1, \dots, L$ , and  $X_2 \stackrel{\mathcal{L}}{=} \sigma Z + \mu$ .

**Proof** See the Appendix in the Supplementary Materials. □

We also see from Proposition (5) that the stochastic representation

$$Y \stackrel{\mathcal{L}}{=} Y_1 + Y_2$$

holds irrespectively of the number of mixture components.

For concreteness, we show the results obtained from Proposition (5) in the case of normal mixtures. Numerical computation of the coefficients  $(|\varphi_Y(k)|)_{k \geq 1}$  for  $k \leq 5$  and selected pairs  $(\mu_1, \sigma_1)$  and  $(\mu_2, \sigma_2)$  is provided in Table 4. The chosen parameters are such that the means belong to  $[0.5, 3.5]$  and the standard deviations are in  $[1.5, 3.5]$ , but they are representative of a much wider behavior. Their values are reported below, where they are ordered to show increasing departures from the uniform distribution in terms of  $\Delta_\infty$ :

**Table 5** As Table 3, but now for the normal mixtures (a)–(f)

$N$	(a)	(b)	(c)	(d)	(e)	(f)
1	0.0327	0.0276	0.0226	0.0313	0.0433	0.1175
2	0.0363	0.0548	0.0692	0.0716	0.0855	0.1139
3	0.0372	0.0531	0.0692	0.0712	0.0848	0.1207
4	0.0372	0.0544	0.0679	0.0731	0.0867	0.0974
5	0.0372	0.0565	0.0680	0.0735	0.0870	0.1189

(a)  $\mu_1 = 1.0, \mu_2 = 3.5, \sigma_1 = 0.5, \sigma_2 = 1.5$  ( $\Delta_\infty = 0.0372$ )

(b)  $\mu_1 = 1.5, \mu_2 = 5.0, \sigma_1 = 0.5, \sigma_2 = 1.5$  ( $\Delta_\infty = 0.0565$ )

(c)  $\mu_1 = 2.5, \mu_2 = 7.5, \sigma_1 = 1.0, \sigma_2 = 3.0$  ( $\Delta_\infty = 0.0680$ )

(d)  $\mu_1 = 1.5, \mu_2 = 5.0, \sigma_1 = 0.5, \sigma_2 = 2.0$  ( $\Delta_\infty = 0.0735$ )

(e)  $\mu_1 = 1.5, \mu_2 = 4.5, \sigma_1 = 0.5, \sigma_2 = 2.0$  ( $\Delta_\infty = 0.0871$ )

(f)  $\mu_1 = 1.0, \mu_2 = 4.0, \sigma_1 = 0.5, \sigma_2 = 1.0$  ( $\Delta_\infty = 0.1191$ )

For a given parameter combination, the coefficients are decreasing with  $k$ , although not monotonically in this case. Table 5 reports the corresponding values of the Kolmogorov distances  $\Delta_N$  for  $N \leq 5$ .

## 4 Testing the Benford hypothesis

### 4.1 Statement of the problem

Working with representation (2), a straightforward translation of the Benford hypothesis into a null hypothesis to be tested is

$$H_0 : s(X) \stackrel{\mathcal{L}}{=} U, \quad (11)$$

where  $U$  is a uniform random variable on  $[0, 1[$ . Given a random sample of  $n$  observations from  $X$ , say  $(X_1, \dots, X_n)$ , several statistics already exist for testing (11). The simplest and possibly most popular one in applications is the first-digit Pearson statistic

$$\chi^2 = \sum_{d=1}^9 \frac{(n\hat{p}_d - np_d)^2}{np_d}, \quad (12)$$

where, for  $d \in \{1, 2, \dots, 9\}$ ,

$$p_d = \log_{10} \left( \frac{d+1}{d} \right) \quad (13)$$

is the theoretical first-digit probability for digit  $d$  when  $H_0$  is true, while

$$\widehat{p}_d = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{d\}}(D_1(X_i))$$

is its sample estimate, based on the first significant digit function  $D_1(x) = \lfloor S(x) \rfloor$ , and  $\mathbf{1}_B$  denotes the indicator function of the set  $B$ .

Strictly speaking,  $\chi^2$  would test the hypothesis that the first-digit distribution of  $X$  conforms to (13), which is a weaker hypothesis than the Benford one. Indeed, the Benford hypothesis (11) on the significant implies that the first digit is Benford, i.e., that expression (13) holds, even if (13) does not imply the significant to be Benford (see, e.g., Berger and Hill 2015). We refer to Barabesi et al. (2022) for a detailed study of the relationships between (11), (13) and another amazing feature of Benford’s law known as the (first-digit) sum-invariance property. However, since disagreement with (13) implies rejection of (11),  $\chi^2$  is also a legitimate test of the Benford hypothesis, provided that its null distribution is derived under  $H_0$ .

A popular competitor of  $\chi^2$  is the mean absolute deviation of first-digit proportions and Benford’s probabilities

$$M = \frac{1}{9} \sum_{d=1}^9 |\widehat{p}_d - p_d|,$$

which is usually called the MAD. A more or less formal use of this statistic is often advocated for anti-fraud purposes in financial applications (Nigrini 2012; Barney and Schulzke 2016), perhaps after extension to the second digit, on the ground that this statistic gives the same weight to all the digits under consideration. A test statistic specifically tailored to (11) is instead the Kolmogorov–Smirnov (two-sided) statistic defined as

$$K = \sup_{t \in [1,10]} |\widehat{F}_{S(X)}(t) - \log_{10} t|, \tag{14}$$

where  $\widehat{F}_{S(X)}$  is the empirical distribution function of  $(S(X_1), \dots, S(X_n))$ .

The test statistics considered above, as well as other related test statistics, suffer from two main drawbacks. First, most of them are derived under specific and non-equivalent properties of Benford’s law, such as the first-digit behavior (13) which underlies both  $\chi^2$  and  $M$ , or the sum-invariance property which motivates the Hotelling-type statistic described in the Supplementary Materials. Therefore, as such, they can show excellent performance under certain circumstances but not in general, as shown by Barabesi et al. (2022). Second, as anticipated in Sect. 1, they are not scale invariant. A scale invariant version of  $K$  is the Kuiper statistic which is computed as

$$V = \left\{ \max_{1 \leq i \leq n} \left( \frac{1}{n} - S(i) \right) + \max_{1 \leq i \leq n} \left( S(i) - \frac{i-1}{n} \right) \right\}, \tag{15}$$

where  $S_{(i)} < \dots < S_{(n)}$  are the order statistics of  $(S(X_1), \dots, S(X_n))$ . However, a serious drawback of  $V$  is its loss of power under multimodal alternatives (Pycke 2010).

## 4.2 Rayleigh-type tests

Examination of (6) inspires the simple strategy of assessing whether a monotone function of the coefficients  $(|\varphi_Y(k)|)_{k \geq 1}$  is non-null. An off-the-shelf option for testing  $H_0$  is the well-known Rayleigh statistic for directional data, which in our problem takes the form  $Z_{1,n}^2 = 2n|\widehat{\varphi}_Y(1)|^2$ , where  $\widehat{\varphi}_Y(k) = \bar{C}_{k,n} + i\bar{S}_{k,n}$ , with

$$\bar{C}_{k,n} = \frac{1}{n} \sum_{i=1}^n \cos(2\pi ks(X_i)), \quad \bar{S}_{k,n} = \frac{1}{n} \sum_{i=1}^n \sin(2\pi ks(X_i)).$$

Unfortunately, this test validates antipodal symmetry, which features both uniform shapes and several multimodal structures. It can thus be safely used only against unimodal alternatives. For a given  $N$ , Buccheri and De Jager (1989) introduced the statistic

$$Z_{N,n}^2 = 2n \sum_{k=1}^N (\bar{C}_{k,n}^2 + \bar{S}_{k,n}^2),$$

which is effective against more general alternative hypotheses. Such a test statistic has been investigated by Bogdan et al. (2002) in a parametric framework for directional data, showing that it is asymptotically optimal for testing uniformity within the family of circular exponential distributions. Data-driven criteria for the selection of  $N$  are considered in Sect. 4.4. Before then, we provide a firm foundation to  $Z_{N,n}^2$  by showing that it is the score statistic for testing (11) with a fixed  $N$ .

## 4.3 Likelihood ratio test of the Benford hypothesis

As we have seen in Sect. 3,  $G_N$  proves to be a suitable approximation of (5) even for small  $N$ . Thus, a general model for the probability density function of  $s(X)$  may be obtained by differentiating (6). We denote this function by  $g = G'_N$ , where the subscript  $N$  is suppressed for simplicity of notation. Reparametrizing, we thus have

$$g(u) = 1 + 2 \sum_{k=1}^N (a_k \cos(2\pi ku) + b_k \sin(2\pi ku)), \quad u \in [0, 1], \quad (16)$$

where  $a_k = |\varphi_Y(k)| \cos(\vartheta_k)$  and  $b_k = |\varphi_Y(k)| \sin(\vartheta_k)$ , for  $k = 1, 2, \dots, N$ .

In order to achieve a nonnegative trigonometric polynomial in (16), i.e., a bona fide probability density function, the Fejér–Riesz condition must hold. More precisely, expression (16) is nonnegative for each  $u \in [0, 1]$  if there exists a vector of order



$(N + 1)$  of complex numbers, say  $c = (c_0, c_1, \dots, c_N)^T$ , such that for  $k = 1, \dots, N$

$$a_k - ib_k = \sum_{j=0}^{N-k} c_{j+k} c_j^*,$$

with  $\sum_{k=0}^N |c_k|^2 = 1$  and where  $c_k^*$  denotes the complex conjugate of  $c_k$ . Model (16) corresponds to a large family of distributions first introduced by Fernández-Durán (2004) in the framework of directional statistics. On the basis of the Fejér–Riesz Theorem (Grenander and Szegő 1984, Section 1.12),  $g(u)$  can be written in a different parametrization as

$$g(u) = c^* T_u c, \quad u \in [0, 1], \tag{17}$$

where  $c^*$  is the conjugate transpose of vector  $c$ , while  $T_u = (t_{k,j,u})$  is a Toeplitz matrix of order  $(N + 1)$  such that  $t_{k,j,u} = e^{2\pi i(k-j)u}$ . The parameter space is assumed to be  $\mathcal{C} = \{c \in \mathbb{C}^{N+1} : c_0 \in \mathbb{R}, c^* c = 1\}$ . Hence, there exist  $(2N + 2)$  parameters in model (17), even if solely  $2N$  of them are free owing to the two constraints. Under model (17), the Benford hypothesis becomes equivalent to

$$H_0 : c_k = 0, \quad k = 1, \dots, N, \tag{18}$$

for a given  $N$ . The likelihood ratio test statistic for assessing (18) is

$$R_{N,n} = \prod_{i=1}^n (\hat{c}^* T_{s(X_i)} \hat{c})^{-1},$$

where  $\hat{c}$  is the maximum likelihood estimator of  $c$ , which can be obtained by means of the algorithm proposed by Fernández-Durán and Gregorio-Domínguez (2010). In addition, the null finite-sample distribution of  $R_{N,n}$  can be approximated through a simple but computationally efficient Monte Carlo algorithm, by bearing in mind that the random variable

$$X = 10^U \tag{19}$$

satisfies the Benford property (1) and that  $S(10^U) = 10^U$ . The usual null large-sample approximation  $\Lambda_{N,n} = -2 \log R_{N,n} \xrightarrow{\mathcal{L}} \chi_{2N}^2$  is also available as  $n \rightarrow \infty$ .

The following proposition states the large-sample equivalence of  $\Lambda_{N,n}$  and  $Z_{N,n}^2$ . Its proof proceeds by showing that  $Z_{N,n}^2$  is the score test statistic for assessing  $H_0$  under model (17). Therefore,  $\Lambda_{N,n}$  and  $Z_{N,n}^2$  share the same optimal properties in the large-sample setting.

**Proposition 6** Under  $H_0$ ,  $\Lambda_{N,n} = Z_{N,n}^2 + o_P(1)$ .

**Proof** See the Appendix in the Supplementary Materials. □

#### 4.4 Estimating the number of trigonometric components

Proposition 6 allows us to extend the asymptotic properties of  $\Lambda_{N,n}$  to  $Z_{\hat{N},n}^2$ , and vice versa. Model selection principles have been mainly addressed from the point of view of the score test. If a simplified version of the BIC criterion is used (see Kallenberg and Ledwina 1995),  $N$  is chosen as the smallest integer for which

$$\psi(N) = Z_{\hat{N},n}^2 - 2N \log(n) \quad (20)$$

is maximum. Therefore, the estimated number of trigonometric components in a data-driven version of  $Z_{\hat{N},n}^2$  is

$$\hat{N} = \inf\{N \in \mathbb{N} : \psi(N) \geq \psi(M), M \in \mathbb{N}\}. \quad (21)$$

It follows from the results of Bogdan et al. (2002) that  $P(\hat{N} = \infty) = 0$  and that, under (18),  $\hat{N} \xrightarrow{P} 1$  as  $n \rightarrow \infty$ , while the test based on  $Z_{\hat{N},n}^2$  is consistent against every alternative to uniformity. In addition,  $Z_{\hat{N},n}^2 \xrightarrow{\mathcal{L}} \chi_2^2$  as  $n \rightarrow \infty$ , even if convergence has been shown to be slow by Bogdan et al. (2002). In Sect. 5 we thus obtain Monte Carlo estimates of the null critical values of  $Z_{\hat{N},n}^2$  through relationship (19).

The asymptotic equivalence of the two tests shown in Proposition 6 ensures that the results on data-driven selection of  $N$  for  $Z_{\hat{N},n}^2$  hold for  $\Lambda_{N,n}$  as well. As a consequence, selection rule (21) is still valid when  $\Lambda_{\hat{N},n}$  is the test statistic of choice if we take

$$\psi(N) = \Lambda_{N,n} - 2N \log(n) \quad (22)$$

as our objective function. The proper BIC criterion is adopted in this case.

#### 4.5 Scale invariance

A desirable property of a test statistic  $T = T(X_1, \dots, X_n)$  is its invariance under scale transformation of the data, which requires that

$$P(T(X_1, \dots, X_n) = T(\sigma X_1, \dots, \sigma X_n)) = 1,$$

for  $\sigma \in \mathbb{R}^+$ . Given a conversion from parsecs to light-years (as in the application of Sect. 6), or from decimeters to feet (as in the application of Sect. 7), while the observation values may change, any inferential statement about the Benford hypothesis should remain unaffected. However, most of the available test statistics are based on functions of the first significant digits and do not share this property.

It is well known that significant digits are scale invariant if and only if the underlying random variable  $X$  is Benford (Berger and Hill 2015, Section 5.1). But even when the Benford hypothesis holds, if we consider for instance the first-digit chi-square statistic  $\chi^2 = \chi^2(X_1, \dots, X_n)$ , it only follows that  $\chi^2(X_1, \dots, X_n) \stackrel{\mathcal{L}}{=} \chi^2(\sigma X_1, \dots, \sigma X_n)$ .

**Table 6** Estimated 0.95th quantiles under the Benford hypothesis for different sample sizes and  $B = 10^6$  replicates

$n$	$Z_{1,n}^2$	$Z_{2,n}^2$	$Z_{\widehat{N},n}^2$	$A_{1,n}$	$A_{2,n}$	$A_{\widehat{N},n}$
50	6.0	9.4	6.9 (1.020)	6.1	9.9	7.5 (1.026)
100	6.0	9.4	6.4 (1.010)	6.0	9.6	6.5 (1.011)
200	6.0	9.5	6.2 (1.006)	6.0	9.6	6.2 (1.007)
500	6.0	9.5	6.1 (1.002)	6.0	9.5	6.1 (1.002)
$\infty$	5.991	9.488	5.991	5.991	9.488	5.991

For  $Z_{\widehat{N},n}^2$  and  $A_{\widehat{N},n}$ , the estimate of  $E[\widehat{N}]$  is reported within parentheses

On the other hand, both the score and the likelihood ratio tests of the Benford hypothesis satisfy the following:

**Proposition 7** *The statistics  $Z_{N,n}^2$  and  $R_{N,n}$  are scale invariant.*

**Proof** See the Appendix in the Supplementary Materials. □

## 5 Simulation experiments

In this section we provide simulation evidence of the empirical advantages of the tests derived in Sect. 4. We compare the performance of the score and likelihood ratio tests—both with  $N \in \{1, 2\}$  and  $N$  selected from the data—to those of  $\chi^2$  and  $M$ , taken as representatives of first-digit tests, the Kolmogorov–Smirnov test (14), perhaps the most obvious choice for assessing the uniformity hypothesis (11), and the scale-invariant Kuiper test (15). For simplicity we relegate to the Supplementary Materials the results for two other tests that we have taken into account in our experiments, the first-digit Hotelling-type test of Barabesi et al. (2022) and the two-digit version of  $\chi^2$ , but we can summarize that comparison with the score and likelihood ratio tests yields similar qualitative findings. We only display power results for  $n = 100$ , while those for different sample sizes are reported in the Supplementary Materials, together with further details about the simulation algorithm.

### 5.1 Exact null distributions

In our first simulation,  $B$  Monte Carlo replicates of the test statistic  $T = T(X_1, \dots, X_n)$  are generated under the Benford hypothesis as

$$T_b^\dagger = T(10^{U_{b,1}}, \dots, 10^{U_{b,n}}), \quad b = 1, \dots, B, \tag{23}$$

where  $U_{b,1}, \dots, U_{b,n}$  are independent uniform random variables on  $[0, 1[$ . For a realization  $t$  of  $T$ , the exact  $p$ -values are computed as  $p_T^\dagger(t) = 1 - F_T^\dagger(t)$ , where  $F_T^\dagger(t) = \frac{1}{B} \sum_{b=1}^B \mathbf{1}_{]-\infty, t]}(T_b^\dagger)$ . In Table 6 we show the estimated  $(1 - \gamma)$ th quantile

$$t_{T,1-\gamma}^{\dagger} = \inf\{t \in \mathbb{R} : F_T^{\dagger}(t) \geq 1 - \gamma\}$$

for each of the test statistics described in Sects. 4.2–4.4, for  $\gamma = 0.05$  and using  $B = 10^6$  replicates under the Benford hypothesis. We also compare the estimated exact quantiles to their asymptotic counterparts. It is clearly seen that asymptotic quantiles are already very accurate with  $n = 50$  for both  $Z_{N,n}^2$  and  $\Lambda_{N,n}$  if  $N$  is fixed in advance, a result that parallels those of Cerioli et al. (2019) and Barabesi et al. (2022) for  $\chi^2$  and other first-digit statistics. Convergence is instead much slower when  $N$  is selected from the data. The estimates  $\widehat{N}$  exhibit non-negligible positive bias in small samples both under (20) and (22), with respect to the true value  $N = 1$  under the Benford hypothesis. These issues point to the use of the reported exact quantiles in the power comparisons that follow.

## 5.2 Power comparisons

The power of each test under comparison is computed with respect to the corresponding exact 0.95th quantile estimated through (23), with  $B = 10^6$ . We first compare the power, based on  $10^4$  replicates, of the various test statistics under the same models for  $X$  studied in Sect. 3, to highlight agreement with the computed Kolmogorov distances.

In the one-population setting, Table 7 shows that the score and likelihood ratio tests behave very similarly. They are the most powerful ones, except when  $X$  is distributed as  $Be(1/2, 1)$ , a case which is equivalent to the generalized Benford law of parameter  $1/2$  (see Table 9 and the related comments below). The best performance is reached with  $N = 1$ , since the distribution of  $s(X)$  is unimodal under all the considered alternatives and the series (5) is well approximated by the first term, in agreement with the results proved in Sect. 3. Nevertheless, the data-driven choice of  $N$  performs almost equally well if no prior information is available. The power of the Kuiper test, although generally inferior, is not too far from that of the top performers in such unimodal scenarios, while it often improves over the non-invariant tests.

The advantage of  $Z_{\widehat{N},n}^2$  and  $\Lambda_{\widehat{N},n}$  is instead paramount under the two-component mixtures considered in Table 8. There, estimation of the number of components becomes crucial, since both  $Z_{1,n}^2$  and  $\Lambda_{1,n}$  fail to detect the underlying structure, while the best performers  $Z_{2,n}^2$  and  $\Lambda_{2,n}$  require knowledge that the second term dominates in the series (5) (see Table 4). Not surprisingly, also the Kuiper test shows a major gap under most of these two-populations alternatives and the ordering among  $V$ ,  $\chi^2$ ,  $M$  and  $K$  becomes less clear.

Our second power scenario widens the perspective by considering a worst-case instance for the likelihood ratio and score tests. Specifically, we assume that  $X$  is a generalized Benford random variable with parameter  $\alpha$ . In this case we have

$$F_{s(X)}(u) = \begin{cases} u & \alpha = 0 \\ \frac{10^{\alpha u} - 1}{10^{\alpha} - 1} & \alpha \neq 0 \end{cases}$$

for  $u \in [0, 1[$  and  $\alpha \in \mathbb{R}$ . This model has been introduced to represent digit distributions when  $X$  follows a power-law (Pietronero et al. 2001; Barabesi and Pratelli

**Table 7** Estimated power under the one-population models of Sect. 3, when  $n = 100$

$N(\mu, \sigma^2)$											
$\mu$	$\sigma$	$\chi^2$	$M$	$K$	$V$	$Z_{1,n}^2$	$Z_{2,n}^2$	$Z_{\hat{N},n}^2$	$\Lambda_{1,n}$	$\Lambda_{2,n}$	$\Lambda_{\hat{N},n}$
0	1	0.295	0.347	0.251	0.496	0.550	0.444	0.522	0.554	0.448	0.519
0	5	0.355	0.411	0.536	0.494	0.545	0.434	0.514	0.549	0.437	0.511
3	5	0.392	0.414	0.605	0.562	0.611	0.508	0.585	0.616	0.510	0.584
3	10	0.301	0.369	0.274	0.508	0.559	0.447	0.529	0.560	0.450	0.529
5	5	0.537	0.475	0.684	0.798	0.840	0.755	0.821	0.841	0.758	0.820
5	10	0.315	0.398	0.331	0.523	0.578	0.467	0.549	0.580	0.470	0.548
$G(\alpha)$											
$\alpha$	$\chi^2$	$M$	$K$	$V$	$Z_{1,n}^2$	$Z_{2,n}^2$	$Z_{\hat{N},n}^2$	$\Lambda_{1,n}$	$\Lambda_{2,n}$	$\Lambda_{\hat{N},n}$	
1.0	0.063	0.075	0.095	0.095	0.103	0.086	0.097	0.102	0.084	0.094	
1.5	0.100	0.146	0.252	0.235	0.257	0.200	0.234	0.256	0.198	0.234	
2.0	0.200	0.261	0.494	0.485	0.541	0.428	0.513	0.545	0.425	0.508	
3.0	0.714	0.733	0.633	0.940	0.961	0.918	0.953	0.964	0.922	0.956	
4.0	0.987	0.994	0.972	0.999	1.000	0.999	1.000	1.000	0.999	1.000	
$Be(\alpha, \beta)$											
$\alpha$	$\beta$	$\chi^2$	$M$	$K$	$V$	$Z_{1,n}^2$	$Z_{2,n}^2$	$Z_{\hat{N},n}^2$	$\Lambda_{1,n}$	$\Lambda_{2,n}$	$\Lambda_{\hat{N},n}$
0.5	1.0	0.686	0.677	0.860	0.682	0.623	0.611	0.614	0.619	0.605	0.612
1.0	1.0	0.999	0.999	1.000	1.000	0.997	0.998	0.997	0.997	0.998	0.997
0.5	2.0	0.120	0.134	0.126	0.169	0.178	0.144	0.169	0.181	0.145	0.169
1.0	2.0	0.559	0.626	0.683	0.727	0.750	0.683	0.730	0.756	0.692	0.737
1.5	2.0	0.973	0.983	0.996	0.995	0.994	0.993	0.994	0.994	0.994	0.994
$PS(\alpha)$											
$\alpha$	$\chi^2$	$M$	$K$	$V$	$Z_{1,n}^2$	$Z_{2,n}^2$	$Z_{\hat{N},n}^2$	$\Lambda_{1,n}$	$\Lambda_{2,n}$	$\Lambda_{\hat{N},n}$	
0.5	0.047	0.050	0.049	0.056	0.056	0.052	0.054	0.055	0.053	0.054	
0.6	0.103	0.113	0.114	0.126	0.145	0.110	0.134	0.145	0.110	0.132	
0.7	0.445	0.464	0.642	0.626	0.679	0.571	0.652	0.682	0.573	0.650	
0.8	0.994	0.990	0.997	0.999	1.000	0.999	1.000	1.000	0.999	0.999	

The exact test size is  $\gamma = 0.05$

2020) and describes the leading-digit distribution of the sequences of prime numbers and non-trivial Riemann zeta zeroes (Luque and Lacasa 2009). It may also help to smoothly represent departure from the Benford property, which arises in the case  $\alpha = 0$  by continuity.

The new insight provided by this scenario is that it shows the effect of modeling the significand transform  $s(X)$  instead of  $X$  itself, providing harmful alternatives to the Benford hypothesis. Indeed, the probability density function of  $s(X)$  is now monotonically increasing or decreasing on  $[0, 1[$ , according to the sign of  $\alpha$ . The

**Table 8** As Table 7, but now for the two-component normal mixtures  $0.5N(\mu_1, \sigma_1^2) + 0.5N(\mu_2, \sigma_2^2)$ 

Case	$\chi^2$	$M$	$K$	$V$	$Z_{1,n}^2$	$Z_{2,n}^2$	$Z_{\hat{N},n}^2$	$\Lambda_{1,n}$	$\Lambda_{2,n}$	$\Lambda_{\hat{N},n}$
(a)	0.201	0.201	0.104	0.183	0.100	0.373	0.277	0.094	0.371	0.281
(b)	0.665	0.667	0.217	0.531	0.087	0.935	0.894	0.089	0.936	0.903
(c)	0.330	0.367	0.220	0.228	0.073	0.559	0.440	0.082	0.560	0.457
(d)	0.358	0.443	0.252	0.404	0.152	0.746	0.625	0.176	0.749	0.644
(e)	0.262	0.369	0.306	0.353	0.134	0.682	0.562	0.116	0.674	0.570
(f)	0.890	0.861	0.608	0.903	0.690	0.975	0.936	0.644	0.971	0.930

See Sect. 3.5 for the definition of cases (a)–(f)

**Table 9** Estimated power when  $X$  is a generalized Benford random variable of parameter  $\alpha$ , for different values of  $\alpha$  and  $n = 100$ 

$\alpha$	$\chi^2$	$M$	$K$	$V$	$Z_{1,n}^2$	$Z_{2,n}^2$	$Z_{\hat{N},n}^2$	$\Lambda_{1,n}$	$\Lambda_{2,n}$	$\Lambda_{\hat{N},n}$
-1.0	0.997	0.999	1.000	0.999	0.998	0.998	0.998	0.997	0.998	0.997
-0.8	0.954	0.978	0.998	0.985	0.963	0.969	0.965	0.962	0.967	0.963
-0.6	0.712	0.802	0.952	0.850	0.794	0.794	0.791	0.787	0.787	0.783
0.6	0.842	0.839	0.950	0.844	0.783	0.786	0.783	0.780	0.783	0.775
0.8	0.982	0.981	0.998	0.982	0.960	0.967	0.961	0.959	0.964	0.960
1.0	0.999	1.000	1.000	1.000	0.996	0.998	0.997	0.996	0.997	0.997

The exact test size is 0.05

**Table 10** Estimate of  $E[\hat{N}]$  when  $X$  is a generalized Benford random variable of parameter  $\alpha$ , for different values of  $\alpha$  and  $n = 100$ 

	$\alpha = -1.0$	$\alpha = -0.8$	$\alpha = -0.6$	$\alpha = 0.6$	$\alpha = 0.8$	$\alpha = 1.0$
Simplified BIC (20)	1.454	1.236	1.117	1.111	1.238	1.453
Likelihood BIC (22)	1.415	1.228	1.118	1.111	1.223	1.406

generalized Benford model is thus particularly unfavorable to the proposed tests, since several terms are needed to approximate the series (5). Table 9 shows empirical powers under this model for some values of  $\alpha$ , while Table 10 reports the estimate of  $E[\hat{N}]$  for both (20) and (22). It is seen that the performances of the score and likelihood ratio tests are very similar, although the non-invariant statistic  $K$  turns out to be the most powerful solution. As anticipated, a value  $N > 1$  is required to capture the monotone shape of  $f_s(X)$ , a task which is best accomplished by (21) in large samples (see the Supplementary Materials).

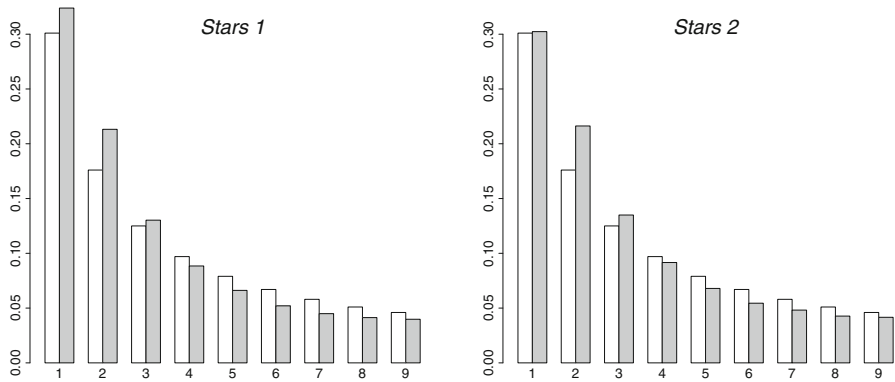
## 6 Astronomical distances

### 6.1 Motivation and data

In our first experiment, we perform two related analyses on the distribution of star distances from Earth, whose original source is the HYG (*Hipparcos*–Yale–Gliese) database available at <https://www.astronexus.com/hyg>. The first investigation is concerned with the data introduced by Alexopoulos and Leontsinis (2014), who explored the star-distance agreement with Benford's law on the basis of the marginal empirical distribution of their first-three significant digits. In our second computation, we improve the analysis by considering a cleaner and more complete version of the same distance database. Our interest in the topic is inspired by several reasons. First, the data are at the center of a quite recent debate in astrophysics and conformance to the Benford hypothesis is for them a still open issue, since there is noticeable uncertainty on the conclusions when comparing empirical evidence with theory. Additionally, previous empirical investigation of this kind of data mostly rely on informal diagnostic checks and single-digit assessment through the non-invariant chi-square test (12). A further reason that motivates our interest is that the validity of the Benford hypothesis has been formally predicted by establishing a link with Hubble's constant for a related problem that involves galaxy distances (Hill and Fox 2016).

Formal statistical inference has not been rigorously conducted on the cited star-distance data in order to validate their fit to Benford's law. Indeed, de Jong et al. (2020) note that the size of astronomical databases is so large that even small departures from the law would constitute an adverse empirical evidence due to huge test power. The HYG 2.0 database contains around 115,000 observations, and this dimension is still very small if compared to the second release of the GAIA database that includes 1.3 billion parallaxes. Therefore, the usual practice has been to calculate a discrepancy index and then judge if it is big (small) enough to informally disprove (support) Benford's law on the basis of its membership to one in a set of subjectively predefined value categories, without any distributional argument. Our suggested strategy for inference on the Benford hypothesis instead relies on bootstrapping the available distance databases for sample sizes comparable to those explored in Sect. 5 and in the Supplementary Materials. A relatively high proportion of rejections of (11) even in these moderately sized random samples would then cast doubt on the Benford nature of the random process generating the observed distances.

The first database that we analyze, and that we name *Stars 1*, consists of the first-three significant digits of star distances to Earth explored by Alexopoulos and Leontsinis (2014) and further examined by Hill and Fox (2016). In the second instance, called *Stars 2*, we replicate a similar experiment on a cleaned version of the HYG 3.0 database, containing 109,398 positive distances recorded in parsecs. Cleaning is performed by removing more than 10,000 stars whose coordinates are deemed to be uncertain and for which distances are replaced by a nonfactual label in the original HYG 3.0 database. In Fig. 1 we include a graphical comparison between theoretical and observed frequencies of the first significant digit in each of the two data sets. Shapes appear to be similar. In both cases the observed data have higher frequencies



**Fig. 1** Left: Comparison of the first significant digit of star distances in database *Stars 1* (grey) with Benford's law (white). Right: The same, but for database *Stars 2*

**Table 11** Astronomical distances. Proportion of rejections of the Benford hypothesis in 1000 samples of size  $n$  drawn with replacement from each database, at exact test size  $\gamma$

$n$	$\gamma$	Stars 1		Stars 2	
		$\chi^2$	$\Lambda_{1,n}$	$\chi^2$	$\Lambda_{1,n}$
200	0.10	0.332	0.606	0.291	0.587
300	0.05	0.337	0.673	0.275	0.650
500	0.01	0.330	0.726	0.268	0.686

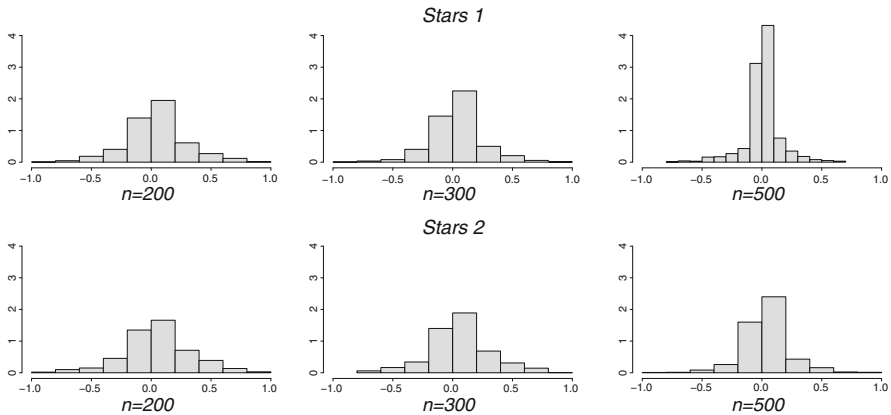
for the first three digits, and smaller for the others, perhaps suggesting a convolution between a noise law and the Benford one. This would call for a cleaner and possibly larger database, as suggested by de Jong et al. (2020).

## 6.2 Results

In our analysis we draw 1000 samples of sizes 200, 300 and 500 with replacement from both star-distance databases. All the generated distance samples can be accessed through the Supplementary Materials. We then compute the proportion of rejections of the Benford hypothesis in each scenario, at exact size  $\gamma$ , for all the tests considered in Sect. 5. Detailed results are given in the Supplementary Materials, while Table 11 focuses on the comparison between  $\chi^2$  and  $\Lambda_{1,n}$  for a few representative values of  $\gamma$ . The score statistic  $Z_{1,n}^2$  and the data-driven selection of  $N$  available through (20) and (22) are seen to yield similar conclusions. There is considerable evidence against the Benford hypothesis, with rejection proportions systematically much larger than  $\gamma$  and with  $\Lambda_{1,n}$  showing to be even less ambiguous than its competitors. Our likelihood ratio approach thus advises that the qualitative assessment of Alexopoulos and Leontsinis (2014) might have been overoptimistic, when contrasted to the full probabilistic structure of  $s(X)$  implied by (11).

We argue that one possible explanation of the disagreement between our results and optimistic expectations about the Benford behavior of star distances may reside in the actual quality of available data. While in the Supplementary Material we show





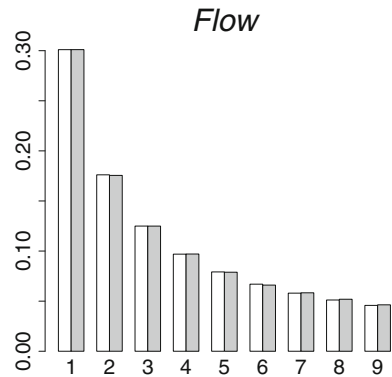
**Fig. 2** Astronomical distances. Distributions of the differences between the  $p$ -values of  $\chi^2$  resulting from data expressed first in parsecs and then in light years

that truncation of  $X$  to three significant digits, as is the case in database *Stars 1*, should not alter conclusions about the Benford hypothesis with the sample sizes under consideration, we note that the data-cleaning step operated on the distances of *Stars 2* slightly reduces the disagreement with (11). More importantly, many ties appear in the HYG database and are then reflected in the realized values of  $s(X)$ , even when several digits are considered. These ties should not occur under Benford's law and their effect is magnified when looking at the empirical distribution of  $s(X)$ . Therefore, the current levels of measurement precision and the associated errors that affect databases such as HYG seem to hinder formal statistical support to the sophisticated mathematical arguments put forth by Hill and Fox (2016).

A similar conclusion may also be reached by considering the measurement error model used in the *Hipparcos* and *GAIA* data processing (Bailer-Jones 2015). For a star at true distance  $r$ , its true but unknown parallax is  $1/r$ . The measured parallax  $\varpi$  is a noisy measurement of  $1/r$  under the small angle approximation. It is assumed that  $\varpi$  is normally distributed with unknown mean  $1/r$  and known standard deviation  $\sigma_\varpi$ , which ultimately depends on the inverse of the number of photons received from the star. It is then hard to anticipate conformance to the Benford hypothesis in view of the results of Sect. 3. Correspondingly, such a measurement-error model may lead to the simulation of synthetic astronomic data, under the Benford hypothesis, as the convolution of a Benford random variable with a suitable noise random variable.

We conclude our analysis of astronomical distances by briefly addressing the issue of lack of scale invariance of standard tests when applied to star-distance data. It is well known that parsecs and light years are two similarly popular scales for these distances. The results for  $\chi^2$  in Table 11 refer to parsecs. Figure 2 shows the distributions of the differences between the  $p$ -values for  $\chi^2$  that are obtained when the values in each sample are rescaled to measure distances in light years. We observe bell shapes roughly centered on zero and an ubiquitous remarkable variability. This confirms that the adoption of a different measurement unit can lead to strongly different results with digit statistics, a hardly acceptable outcome in the scientific context under consideration.

**Fig. 3** Comparison of the first significant digit of the streamflow data (grey) with Benford's law (white)



## 7 Streamflow data

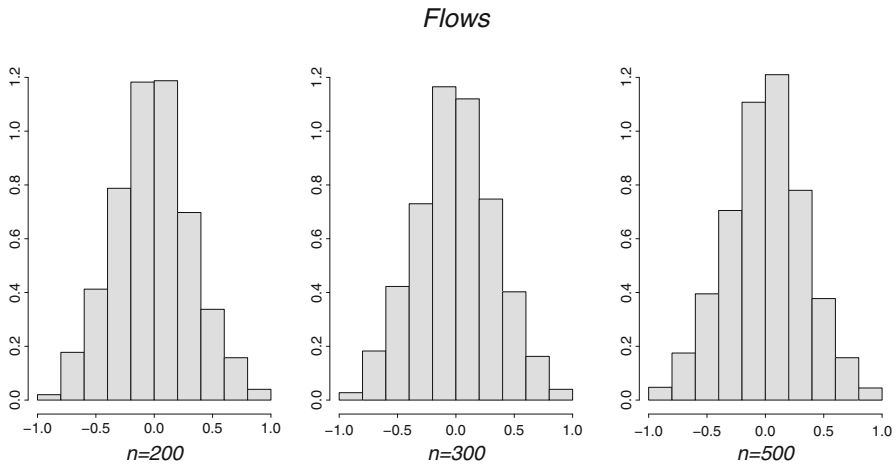
Inspired by a Referee, we present a second case study related to water flows (in cubic feet per second) recorded at more than 17,000 streamgauge sites in the U.S. over an extended period of time spanning from 1874 to 2004. This quite extensive collection, that contains 457,440 observations, was obtained and analyzed by Nigrini and Miller (2007) after deletion of blanks, zeroes (possibly corresponding to absence of water flows), negative numbers (possibly corresponding to data errors) and duplicate records. Such a large hydrological data set comes with the R package *benford.analysis* available at the web address <https://github.com/carloscinelli/benford.analysis>, from which we have retrieved it (see also Cinelli 2022).

Nigrini and Miller (2007) and Nigrini (2012, Chapter 12), through a detailed data description and a battery of formal digit tests, conclude that the streamflow data set exhibits a practically excellent fit to Benford's law, in spite of the possible "excess of power" related to the big sample size and in spite of some potential sampling bias in the spatial distribution of the measuring stations. Based on first digit counts reported in Fig. 3, we easily agree since the visual fit is nearly perfect. Therefore, differently from the astronomical case study, our aim here is to inspect how our findings apply when the Benford hypothesis is a reasonable approximation to the true digit generating process.

As a first experiment, we test the whole data set, obtaining  $p$ -values equal to 0.0342 and 0.0144 for  $\chi^2$  and  $\Lambda_{1,n}$ , respectively. Although such quantities may appear smaller than expected, they are in good agreement with the evidence provided by Nigrini and Miller (2007). We thus see that also our likelihood ratio approach does not lead to strong evidence against the Benford hypothesis in spite of the huge available sample and of possible inaccuracies in the data collection process. By way of comparison, we note that for the astronomical distance data both tests yield a virtually null  $p$ -value, even with a smaller sample size. We then repeat the sub-sampling procedure of Sect. 6 to streamflow data. Again, we see in Table 12 that both tests exhibit rejection proportions which are very close to nominal test sizes, thus further corroborating the plausibility of the Benford hypothesis in this problem.

**Table 12** Streamflow data. Proportion of rejections of the Benford hypothesis in 2000 samples of size  $n$  drawn with replacement from the database, at exact test size  $\gamma$

$n$	$\gamma$	$\chi^2$	$A_{1,n}$
200	0.10	0.090	0.104
300	0.05	0.044	0.045
500	0.01	0.009	0.010



**Fig. 4** Streamflow data. Distributions of the differences between the  $p$ -values of  $\chi^2$  resulting from data expressed first in cubic feet and then in cubic decimeters per second

We conclude our analysis, as in Sect. 6, by assessing the effect of changing the measurement unit of streamflows from cubic feet to cubic decimeters per second. Figure 4 repeats the frame of Fig. 2 by reporting the distributions of the differences between the  $p$ -values of  $\chi^2$  after this change of unit. The results are similar and perhaps even more alarming, meaning that the distributions are clearly less concentrated around zero. This empirical evidence appears to be very interesting because it shows that the sensitivity of  $\chi^2$  to the data scale does not alleviate even if the data are close to being genuinely Benford.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s11749-023-00881-y>.

**Acknowledgements** We are grateful to Editor-in-Chief Prof. Ana María Aguilera, an Associate Editor and two anonymous reviewers for their constructive comments. We also thank Ted Hill for sharing the star-distance data examined in Hill and Fox (2016). Andrea Cerioli has financially been supported by the Programme “AZIONE A BANDO DI ATENEO PER LA RICERCA 2022 dell’Università di Parma: Robust statistical methods for the detection of frauds and anomalies in complex and heterogeneous data”.

**Funding** Open access funding provided by Università degli Studi di Parma within the CRUI-CARE Agreement. Open access funding provided by Università degli Studi di Parma within the CRUI-CARE Agreement.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give

appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Alexopoulos T, Leontsinis S (2014) Benford's law in astronomy. *J Astrophys Astron* 35:639–648
- Álvarez-Esteban PC, del Barrio E, Cuesta-Albertos JA, Matrán C (2012) Similarity of samples and trimming. *Bernoulli* 18:606–634
- Bailer-Jones CAL (2015) Estimating distances from parallaxes. *Publ Astron Soc Pac* 127:994–1009
- Barabesi L, Pratelli L (2020) On the generalized Benford law. *Stat Probab Lett* 160:108702
- Barabesi L, Cerasa A, Cerioli A, Perrotta D (2018) Goodness-of-fit testing for the Newcomb–Benford law with application to the detection of customs fraud. *J Bus Econ Stat* 36:346–358
- Barabesi L, Cerioli A, Perrotta D (2021) Forum on Benford's law and statistical methods for the detection of frauds. *Stat Methods Appl* 30:767–778
- Barabesi L, Cerasa A, Cerioli A, Perrotta D (2022) On characterizations and tests of Benford's law. *J Am Stat Assoc* 117:1887–1903
- Barney BJ, Schulzke KS (2016) Moderating “Cry Wolf” events with excess MAD in Benford's law research and practice. *J Forensic Account Res* 1:A66–A90
- Berger A, Hill TP (2011) Benford's law strikes back: no simple explanation in sight for mathematical gem. *Math Intell* 33:85–91
- Berger A, Hill TP (2015) *An introduction to Benford's law*. Princeton University Press, Princeton
- Berger A, Hill TP (2021) The mathematics of Benford's law: a primer. *Stat Methods Appl* 30:779–795
- Berger A, Twelves I (2018) On the significands of uniform random variables. *J Appl Probab* 55:353–367
- Bogdan M, Bogdan K, Futschik A (2002) A data driven smooth test for circular uniformity. *Ann Inst Stat Math* 54:29–44
- Buccheri R, De Jager O (1989) Detection and description of periodicities in sparse data. suggested solutions to some basic problems. In: Ögelman H, van der Heuvel E (eds) *Timing neutron stars*. Kluwer, Dordrecht, pp 95–111
- Cerasa A (2022) Testing for Benford's law in very small samples: simulation study and a new test proposal. *PLOS ONE* 17(e0271):969
- Cerioli A, Barabesi L, Cerasa A, Menegatti M, Perrotta D (2019) Newcomb–Benford law and the detection of frauds in international trade. *PNAS* 116:106–115
- Cerqueti R, Maggi M (2021) Data validity and statistical conformity with Benford's law. *Chaos Solitons Fractals* 144(110):740
- Cerqueti R, Lupi C (2023) Severe testing of Benford's law. *TEST* <https://doi.org/10.1007/s11749-023-00848-z>
- Cinelli C (2022) Package ‘benford.analysis’. <https://cran.r-project.org/web/packages/benford.analysis/benford.analysis.pdf>. Accessed 13 April 2023
- de Jong J, de Bruijne J, De Ridder J (2020) Benford's law in the Gaia universe. *Astrophys Astron* 642:A205
- del Barrio E, Inouzhe H, Matrán C (2020) On approximate validation of models: a Kolmogorov–Smirnov-based approach. *TEST* 29:938–965
- Demidenko E (2020) *Advanced statistics with applications in R*. Wiley, Hoboken
- Dümbgen L, Leuenberger C (2008) Explicit bounds for the approximation error in Benford's law. *Electron Commun Probab* 13:99–112
- Dümbgen L, Leuenberger C (2015) Explicit error bounds via total variation. In: Miller SJ (ed) *Benford's law: theory and applications*. Princeton University Press, Princeton, pp 119–134
- Engel HA, Leuenberger C (2003) Benford's law for exponential random variables. *Statist Probab Lett* 63:361–365
- Farcomeni A, Punzo A (2020) Robust model-based clustering with mild and gross outliers. *TEST* 29:989–907

- Fernández-Durán JJ (2004) Circular distributions based on nonnegative trigonometric sums. *Biometrics* 60:499–503
- Fernández-Durán JJ, Gregorio-Domínguez M (2010) Maximum likelihood estimation of nonnegative trigonometric sums models using a Newton-like algorithm on manifolds. *Electron J Statistics* 4:1402–1410
- Fernández-Gracia J, Lacasa L (2018) Bipartisanship breakdown, functional networks, and forensic analysis in Spanish 2015 and 2016 national elections. *Complexity* 2018:9684749
- Grenander U, Szegő G (1984) Toeplitz forms and their applications, 2nd edn. Chelsea Publishing Company, New York
- Hennig C (2022) An empirical comparison and characterisation of nine popular clustering methods. *Adv Data Anal Classif* 16:201–229
- Hill TP (1995) A statistical derivation of the significant-digit law. *Stat Sci* 10:354–363
- Hill TP, Fox RF (2016) Hubble’s law implies Benford’s law for distances to galaxies. *J Astrophys Astron* 37:1–8
- Ingrassia S, Jacques J, Yao W (2022) Special issue on “Models and learning for clustering and Classification”. *Adv Data Anal Classif* 16:231–234
- Kallenberg WCM, Ledwina T (1995) Consistency and Monte Carlo simulation of a data driven version of smooth goodness-of-fit tests. *Ann Stat* 23:1594–1608
- Kossofsky AE (2015) Benford’s law: theory, the general law of relative quantities, and forensic fraud detection applications. World Scientific, Singapore
- Lacasa L (2019) Newcomb–Benford law helps customs officers to detect fraud in international trade. *PNAS* 116:11–13
- Leemis L (2015) Benford’s law geometry. In: Miller SJ (ed) Benford’s law: theory and applications. Princeton University Press, Princeton, pp 109–118
- Luque B, Lacasa L (2009) The first-digit frequencies of prime numbers and Riemann zeta zeros. *Proc R Soc A* 465:2197–2216
- Mardia KV, Jupp PE (2000) Directional statistics. Wiley, New York
- Mebane WR Jr (2010) Fraud in the 2009 presidential election in Iran? *Chance* 23:6–15
- Melita MD, Miraglia JE (2021) On the applicability of Benford law to exoplanetary and asteroid data. *New Astron* 89:101654
- Miller SJ (ed) (2015) Benford’s law: theory and applications. Princeton University Press, Princeton
- Miller SJ (2015) Fourier analysis and Benford’s law. In: Miller SJ (ed) Benford’s law: theory and applications. Princeton University Press, Princeton, pp 68–105
- Nigrini MJ (2012) Benford’s Law. Wiley, Hoboken
- Nigrini MJ, Miller SJ (2007) Benford’s law applied to hydrology data—results and relevance to other geophysical data. *Math Geol* 39:469–490
- Olver FWJ, Lozier DW, Boisvert RF, Clark CW (2010) NIST Handbook of mathematical functions. Cambridge University Press, Cambridge
- 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. *Stat Sci* 26:502–516
- Pietronero L, Tosatti E, Tosatti V, Vespignani A (2001) Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf. *Physica A* 293:297–304
- Pinkham RS (1961) On the distribution of first significant digits. *Ann Math Stat* 32:1223–1230
- Pycke JR (2010) Some tests for uniformity of circular distributions powerful against multimodal alternatives. *Can J Stat* 38:80–96
- Shao L, Ma BQ (2010) The significant digit law in statistical physics. *Physica A* 389:3109–3116
- Tam Cho WK, Gaines BJ (2007) Breaking the (Benford) law. *Am Stat* 61:218–223