Computational Statistics (2024) 39:3779–3800
https://doi.org/10.1007/s00180-024-01463-8

ORIGINAL PAPER

Some new invariant sum tests and MAD tests for the assessment of Benford's law

Wolfgang Kössler[1] · Hans-J. Lenz[2] · Xing D. Wang[1]

Received: 25 September 2023 / Accepted: 12 January 2024 / Published online: 13 February 2024

Abstract

The Benford law is used world-wide for detecting non-conformance or data fraud of numerical data. It says that the significand of a data set from the universe is not uniformly, but logarithmically distributed. Especially, the first non-zero digit is One with an approximate probability of 0.3. There are several tests available for testing Benford, the best known are Pearson's χ^2-test, the Kolmogorov–Smirnov test and a modified version of the MAD-test. In the present paper we propose some tests, three of the four invariant sum tests are new and they are motivated by the sum invariance property of the Benford law. Two distance measures are investigated, Euclidean and Mahalanobis distance of the standardized sums to the orign. We use the significands corresponding to the first significant digit as well as the second significant digit, respectively. Moreover, we suggest inproved versions of the MAD-test and obtain critical values that are independent of the sample sizes. For illustration the tests are applied to specifically selected data sets where prior knowledge is available about being or not being Benford. Furthermore we discuss the role of truncation of distributions.

Keywords Benford law · Goodness of fit test · Sum invariance · Data fraud · Data manipulation · Data quality

✉ Wolfgang Kössler
koessler@informatik.hu-berlin.de

Hans-J. Lenz
hans-j.lenz@fu-berlin.de

Xing D. Wang
wangxida@informatik.hu-berlin.de

[1] Institut für Informatik, Humboldt Universität zu Berlin, Rudower Chaussee 25, 12489 Berlin, Germany

[2] Institut für Statistik und Ökonometrie, Freie Universität Berlin, Boltzmannstr. 20, 14195 Berlin, Germany

1 Introduction

In many data sets the first non-zero digit d is not uniformly distributed but obeys a logarithmic law. This fact was observed by Newcomb (1881) and Benford (1938). Conformance officers of big companies use the Benford law for unscrambling data manipulations mostly by applying the χ^2 goodness-of-fit test. Such manipulations may be inserting fraudulent figures or changing digits. Those and more applications may be found, for example, in the books of Nigrini (2012) and Berger and Hill (2015), see also Kössler et al. (2024). However, not every real or artificial data set follows the Benford law, the question arises how this can be tested in practice. There is a vast literature on applications and testing of Benford's law, we refer to the website of Berger et al. (accessed 3.1.2024) and to Nigrini (2012). The latter author applies Pearsons χ^2-test, the Kolmogorov–Smirnov test and the MAD-test on the 1st digit, the 2nd digit, the 1st and 2nd digit together and 1st, 2nd and 3rd digit together. His MAD-test assigns the numerical values of the MAD statistic to the linguistic terms close conformity, acceptable conformity, marginally acceptabel conformity und nonconformity, cf. Table 7.1 (p. 160) of Nigrini (2012) book.

Berger and Hill (2011) as well as Nigrini (1992) analyzed the scale-, base- and sum-invariance. The latter includes especially that the expected sum of all the significands with leading digit 1 is equal to the sums of the significands of the remaining digits 2, ..., 9, respectively.

In the present article we apply the sum invariance properties of Benford's law for constructing several further tests of significance. Our test statistics are suitable linear combinations of squares of suitably chosen statistics, and they are, under the null hypothesis, asymptotically or approximately χ^2-distributed.

Emphasis is especially taken on the second significant digit. A χ^2 goodness-of-fit test for the second digit was already suggested, cf. eg. Diekmann (2007). We suggest some further tests based on properties of the second significant digit. In Sect. 2.1 we present some basic properties and some statistical tests for testing Benford that are applied later on. In Sect. 2.2 we recall the χ^2 goodness-of-fit test, the Kolmogorov–Smirnov test, and apply them to the first and second significant digit. Moreover, the MAD-test is modified to obtain critical values that do not depend on the sample size. In Sect. 2.3 we introduce four variants of the invariant sum test, three of them are new, and in Sect. 3 we illustrate the considered tests on some chosen data sets. In Sect. 4 we summarize and discuss the results. All mathematical derivations are deferred to the appendices.

2 Methodology

2.1 Some basics of the Benford law

Benford's law makes claims about the leading digits of a number regardless of its scale. Closely connected to the leading digits are the terms of significands and significant digits, which formal notion is given in Definition 1.

Definition 1 (Significant digits and the significand, Berger and Hill (2015)) Let $x \in \mathbb{R}$. The first significant digit $D_1(x) = d$ of x is given by the unique integer $d \in \{1, 2, \ldots, 9\}$ where $10^k d \leq |x| < 10^k(d + 1)$ with an integer k.

The m-th significant digit $D_m(x) = d$ with $m \geq 2$ can recursively be determined by $10^k\left(\sum_{i=1}^{m-1} D_i(x)10^{m-i} + d\right) \leq |x| < 10^k\left(\sum_{i=1}^{m-1} D_i(x)10^{m-i} + d + 1\right)$ where $d \in \{0, 1, \ldots, 9\}$ and $k \in \mathbb{Z}$.

The significand function $S : \mathbb{R} \rightarrow [1, 10)$ is defined as follows: If $x \neq 0$ then $S(x) = t$, where t is the unique number $t \in [1, 10)$ with $|x| = 10^k t$ for some unique $k \in \mathbb{Z}$. For $x = 0$ we set, for convenience, $S(0) := 0$.

Next, we state the strong and weak form of Benford's law.

Definition 2 (Benford's law for the significand, strong form of Benford's law) The significand $S(X)$ follows Benford's law if

$$P(S(X) \leq t) = \log t \text{ for all } t \in [1, 10). \tag{1}$$

Definition 3 (Benford's law for the first significant digit, weak form of Benford's law) The probability of the first significant digit $d \in \{1, 2, 3...9\}$ is $P(D_1(X) = d) = \log(1 + d^{-1})$.

In Table 1, we give the distribution of the leading digit D_1.

In the following we call a random variable X Benford distributed iff (1) is satisfied, and we write $X \sim$ Benford. Benford distributed random variables own some remarkable properties. In the present article we focus on the sum-invariance property. Sum-invariance specifically means that, if summing all significands with the first digit 1 we expect the same sum as summing all significands with the first digit 2, 3 etc., i.e. their expectations are the same. For further explanations and proofs we refer to Berger and Hill (2011, 2015), Pinkham (1961) and Nigrini (1992).

Table 1 Probabilities $P(D_1(X) = d_1)$ according to Benford's law

d_1	1	2	3	4	5	6	7	8	9
$P(d_1)$	0.301	0.176	0.124	0.096	0.079	0.066	0.057	0.051	0.045

2.2 Classical tests against Benford and their modifications

Our test problem is in general

$$H_0 : X \sim \text{ Benford} \qquad \text{against} \qquad H_1 : X \nsim \text{ Benford} .$$

Note, H_1 is a very large class of alternatives.

The χ^2-test is one of the most popular goodness-of-fit tests, and it was originated by Pearson (1900). The χ^2-test statistic measures the relative distance between the relative frequencies n_j/n and the probabilities $p_j = P(D_1 = d_j)$ for all $j = 1, 2, \dots, 9$ under the Benford law, and it is defined by

$$\chi^2 = n \sum_{j=1}^{9} \frac{(n_j/n - p_j)^2}{p_j} = \sum_{j=1}^{9} \frac{(n_j - np_j)^2}{np_j}. \tag{2}$$

The χ^2-test rejects the null hypothesis H_0, if $\chi^2 > \chi^2_{1-\alpha,8}$, where $\chi^2_{1-\alpha,8}$ is the $(1 - \alpha)$ quantile of the χ^2 distribution with eight degrees of freedom. Note that the χ^2 goodness-of-fit test is an approximate test, the statistic (2) is asymptotically χ^2-distributed with eight degrees of freedom.

Since some data fraudsters may know Benford's law for the first significant digit some authors suggest to use the second significant digits instead of the first one and to apply a goodness-of-fit test to them, cf. eg. Diekmann (2007) or Hein et al. (2012) for scientific fraud or Mebane (2010) for election fraud.

The probability of the second significant digit $d \in \{0, 1, 2, \dots, 9\}$ is $P(D_2(X) = d) = \sum_{j=1}^{9} \log_{10}(1 + \frac{1}{10j+d})$, and it is presented in Table 2. Note that according to rounding effects, the probabilities do not exactly sum up to one. We abbreviate both variants of the χ^2 goodness-of-fit test by GoF1 and GoF2, respectively.

An alternative goodness-of-fit test is the Kolmogorov–Smirnov (KS) test, cf. Kolmogorov (1933), Smirnov (1948) and Darling (1957). The idea of this test is to compare the empirical cumulative distribution function (cdf) $F_n(x)$ with a fully specified theoretical one, $F_0(x)$. The KS-tests uses the norm

$$d_{max} = sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)|. \tag{3}$$

Since we investigate tests based on the first or second significant digit, respectively, we apply the KS-test first to the first significant digit according to the weak form of Benford's law (cf. Definition 3). Alternatively, we apply the KS-test to the second significant digit.

Table 2 Probabilities $P(D_2(X) = d_2)$ according to Benford's law

d_2	0	1	2	3	4	5	6	7	8	9
$P(d_2)$	0.1197	0.1139	0.1088	0.1043	0.1003	0.0967	0.0934	0.0904	0.0876	0.0850

Table 3 Asymptotic critical values $c_{KS2,1-\alpha}$ and $c_{MAD2,1-\alpha}$ of the KS2 and MAD2 test

	α		
	0.01	0.05	0.1
KS2	1.46	1.19	1.05
MAD2	3.92	3.42	3.18

Table 4 Critical values $c_{MAD,1-\alpha}$ of the MAD test (1st digit)

	α		
n	0.1	0.05	0.01
72	2.883	3.111	3.618
369	2.896	3.140	3.683
1000	2.905	3.156	3.683
3998	2.895	3.159	3.663
7022	2.881	3.142	3.597
∞	2.869	3.084	3.485

The critical values of the KS test were completely tabulated by Miller (1956) for underlying continuous distributions. Morrow (2014) computed tighter bounds by Monte Carlo simulation for the discrete Benford distribution of the first digit, cf. Table 1.

For the KS-test applied to the second significant digit, cf. Table 2, the (asymptotic) critical values are approximated by simulation. To do this we simulate from a (continuous) Benford distribution (cf. Definition 2). Then we put the observations into bins $0, \ldots, 9$ according to the the definition of the second (significant) digit. Taking a large sample size of $n = 10,000$ and repeating this $M = 10,000$ times we get a sufficiently accurate estimation of the asymptotic critical values. Some critical values are presented in Table 3. We abbreviate the KS-tests, based on the first or second significant digit, respectively, by KS1 and KS2.

As another alternative goodness-of-fit test we suggest a kind of MAD-test that is based on the statistic

$$MAD = \sqrt{n} \sum_{j=1}^{k} |n_j/n - p_j|, \tag{4}$$

where MAD stays for Mean Absolute Deviation. Though there are no means here, our proposal is derived from an idea due to Nigrini (2012) who used the mean $MAD_N = \sum_{j=1}^{k} \frac{|n_j/n-p_j|}{k}$, where the index N stays for Nigrini. Our proposal uses a suitably scaled sum of the absolute deviations between the relative frequencies and the Benford probabilities for the first digit. In our new version we introduced the factor $\sqrt{n}$ to get critical values of the test being rather independent of n. This property is illustrated in Table 4. The motivation for introducing the factor $\sqrt{n}$ is that the relative frequencies tend to the true probability with $\sqrt{n}$ rate. Recently, Cerqueti and Lupi (2021) obtained the asymptotic distribution of the MAD statistic (4). From that we computed the asymptotic critical values, cf. Table 4. The convergence of the finite critical values to the asymptotic critical value is rather fast.

Evidently, the critical values are not very sensitive to the sample sizes. For simplicity, we use in our study the critical value $c_{MAD,1-\alpha} = 3.60$ for $\alpha = 0.01$.

The MAD-test may also be applied to the second significant digit. Some (asymptotic) critical values are presented in Table 3. The critical values are obtained in an analogous way as that for the tests KS1 and KS2. We abbreviate both variants, first and second significant digit, by MAD1 and MAD2.

One may ask why we do not use the first two digits together. This idea was suggested in, Diekmann (2007) cf. also Nigrini (2012). However, we consider data sets with moderate sample sizes, nearly between n = 200 and n = 4000. If we use the first two digits together then we have altogether 90 bins and therefore very much bins with very few or even no observations. Therefore this idea is not applicable here to the class of invariant sum tests. However, some kind of KS-test or MAD-test for discrete distributions might be applied. Since our interest here lies on invariant sum tests, they are not considered.

Of course, there are other possibilities to test against Benford, despite of the invariant sum tests that we introduce in the next section. We mention only two recently published ideas. Kazemitabar and Kazemitabar (2022) make use of the alternative definition of Benford's Law saying that the logarithms of the significands are uniformly distributed. Cerqueti and Maggi (2021) discuss some distance measures, especially the sum of squares deviation and the MAD.

2.3 Invariant sum tests

In this section we apply the invariant-sum property of Benford, cf. Nigrini (1992), Allaart (1997) and Berger and Hill (2015, theorem 5.18).Berger and Hill (2015) To do this we define the sets $C(d_1,\dots,d_m) = \{x \in [1,10) : D_j(x) = d_j \text{ for } j = 1,\dots,m\}$, $C_1(d_1) = \{x \in [1,10) : D_1(x) = d_1\} = [d_1, d_1 + 1)$ and $C_2(d_2) = \{x \in [1,10)$: $D_2(x) = d_2\} = \bigcup_{j=1}^{9} [j + \frac{d_2}{10}, j + \frac{d_2+1}{10})$. $C(d_1,\dots,d_m)$ is the set of all significands with first m digits $d_1,\dots,d_m$, $C_1(d_1)$ is the set of all significands with first digit d_1, and $C_2(d_2)$ is the set of all significands with second digit d_2.

Proposition 1 (Sum Invariance (Berger and Hill (2015), *Nigrini* (2012), *Allaart* (1997)) *A random variable X is Benford if and only if X has sum invariant significant digits, i.e. for every fixed $m, m \in \mathbb{N}$, the expectations $\mathbb{E}(S(X)\mathbb{1}_{C(d_1,\dots,d_m)}(S(X)))$ are the same for all tuples $(d_1,\dots,d_m), d_1 \neq 0$ of digits.*

Therefore one necessary condition for X to be Benford is that the expectation of the sum of all significands with the first digit 1, 2, 3, ..., 9 is the same. The same is true for the expectation of the sum of all significands with second digit $0, 1, \dots, 9$.

Let us start with the first significant digit. Denote by $\theta = \mathbb{E}\big(S(X)\mathbb{1}_{C_1(i)}(S(X))\big) = \frac{1}{\ln 10}$ the expectation of the random variable

$S(X)\mathbb{1}_{C_1(d_1)}(S(X))$ if Benford is true. Let θ_i be the true expectation of $S(X)\mathbb{1}_{C_1(i)}(S(X))$ for the underlying distribution.

Then our first test problem is

$$H_{0,1} : \theta_1 = \ldots = \theta_9 = \theta \quad \text{against} \quad H_{1,1} : \exists j \in \{1, \ldots, 9\} : \theta_j \neq \theta$$

Denote the sums of the significands of the observations X_i in the interval $[j, j+1)$

$$\text{Sum}_{1,j} = \sum_{i=1}^{n} S(X_i)\mathbb{1}_{C_1(j)}(S(X_i)).$$

Since we have sums of n independent identically distributed random variables $S(X_i)\mathbb{1}_{C_1(j)}(S(X_i)), i = 1, \ldots, n$, and they have finite variance, we may assume that they are approximately normally distributed, and the standardized sums

$$R_{1,j} = \frac{\text{Sum}_{1,j} - \mathbb{E}(\text{ Sum}_{1,j})}{\sqrt{\text{var }(Sum_{1,j})}}$$

are (approximately) standard normal. The expectations $\mathbb{E}(\text{ Sum}_{1,j}) = \frac{n}{\ln 10}$, variances $\text{var}(\text{Sum}_{1,j})$ and covariances are derived in the Appendix A.

Let be $\mathbf{R}_1 = (R_{1,1}, \ldots, R_{1,9})$ and $\mathbf{\Sigma}_{R_1}$ be the correlation matrix of the vector $\mathbf{R}_1$of standardized sums under the null. We consider the following two types of test statistics

$$IS_{1,E} = \mathbf{R}_1'\mathbf{R}_1 \qquad \text{and} \qquad IS_{1,M} = \mathbf{R}_1'\mathbf{\Sigma}_{R_1}^{-1}\mathbf{R}_1$$

where IS stays for **I**nvariant **S**um. The statistic $IS_{1,E}$ is the Euklidean distance of the vector $\mathbf{R}_1$ of standardized sums from zero, and $IS_{1,M}$ is the corresponding Mahalanobis distance.

The question may come up why we use both distance measures, Euclidean and Mahalanobis. The two distances are generally different, and so are the corresponding test statistics. Therefore there may be alternative directions for which the Euclidean distance is better than the Mahalanobis distance and vice versa.

Theorem 2 *Under $H_{0,1}$ the statistic $IS_{1,M}$ is asymptotically χ^2-distributed with nine degrees of freedom, and $IS_{1,E}$ is is approximated by a weighted sum of independent χ^2-distributed random variables, each with one degree of freedom.*

Table 5 Simulated levels of significance under $H_{0,1}$ and $H_{0,2}$, respectively, of the invariant sum tests, for various sample sizes, nominal level of significance $\alpha = 0.01$

n	25	100	400	900
$IS_{1,E}$	0.012	0.010	0.011	0.011
$IS_{1,M}$	0.011	0.009	0.011	0.011
$IS_{2,E}$	0.023	0.012	0.015	0.010
$IS_{2,M}$	0.023	0.011	0.013	0.009

The proof of the theorem can be found in Appendix B.

The null hypothesis $H_{0,1}$ is rejected in favour of $H_{1,1}$ if $IS_{1,M} > \chi^2_{1-\alpha,9}$ or if $IS_{1,E} > c_{IS_{1,E},1-\alpha}$, respectively, where $\chi^2_{1-\alpha,9}$ is the $1-\alpha$-quantile of the χ^2-distribution with nine degrees of freedom and $c_{IS_{1,E},1-\alpha}$ is the corresponding quantile of the null distribution of $IS_{1,E}$. The latter quantile will be determined by approximating the null distribution of $IS_{1,E}$ by a suitably scaled and shifted χ^2-distribution, see Appendix C. Table 5 gives simulated levels of significance of the two tests. Even for small sample sizes they are close to the nominal value of $\alpha = 0.01$.

Note that statistic $IS_{1,M}$ was independently introduced by Barabesi et al. (2021).

Now, consider the second significant digit. Denote by $\vartheta = \mathbb{E}\big(S(X)\mathbb{1}_{C_2(j)}(S(X))\big) = \frac{9}{10\ln 10}$ the expectation of $S(X)\mathbb{1}_{C_2(j)}(S(X))$ if Benford is true. Let ϑ_j the true expectation of $S(X)\mathbb{1}_{C_2(j)}(S(X))$ for the underlying distribution.

Then our second test problem is

$$H_{0,2}\,:\,\vartheta_0 = \ldots = \vartheta_9 = \vartheta \quad \text{against} \quad H_{1,2}\,:\,\exists j \in \{0,\ldots,9\}\,:\,\vartheta_j \neq \vartheta$$

Denote the sums of the significands in $C_2(j)$ of observations X_i

$$\mathrm{Sum}_{2,j} = \sum_{i=1}^{n} S(X_i)\mathbb{1}_{C_2(j)}(S(X_i)).$$

Again, we have sums of n independent identically distributed random variables $S(X_i)\mathbb{1}_{C_2(j)}(S(X_i)), i = 1,\ldots,n$, and they have finite variance, we may assume that they are approximately normally distributed, and the standardized sums

$$R_{2,j} = \frac{\mathrm{Sum}_{2,j} - \mathbb{E}(\,\mathrm{Sum}_{2,j})}{\sqrt{\mathrm{var}(\mathrm{Sum}_{2,j})}}$$

are (approximately) standard normal. The expectations $\mathbb{E}(\,\mathrm{Sum}_{2,j}) = \frac{9n}{10\ln 10}$, variances $\mathrm{var}(\mathrm{Sum}_{2,j})$ and covariances are derived in the Appendix A.

Let be $\mathbf{R}_2 = (R_{2,0},\ldots,R_{2,9})$ and let $\mathbf{\Sigma}_{R_2}$ be the correlation matrix of the sums vector $\mathbf{R}_2$ under the null. Similarly as above we consider the following two types of test statistics

$$IS_{2,E} = \mathbf{R}_2'\mathbf{R}_2 \qquad \text{and} \qquad IS_{2,M} = \mathbf{R}_2'\mathbf{\Sigma}_{R_2}^{-1}\mathbf{R}_2$$

Theorem 3 *Under $H_{0,2}$ the statistic $IS_{2,M}$ is asymptotically χ^2-distributed with ten degrees of freedom, and $IS_{2,E}$ is a weighted sum of independent χ^2-distributed random variables, each with one degree of freedom.*

The proof of the theorem can be found in Appendix B.

Table 5 gives simulated levels of significance of the two tests. Again, even for small sample sizes they are close to the nominal value of $\alpha = 0.01$.

In Appendix F our algorithm for the implementation of the invariant sum tests is provided.

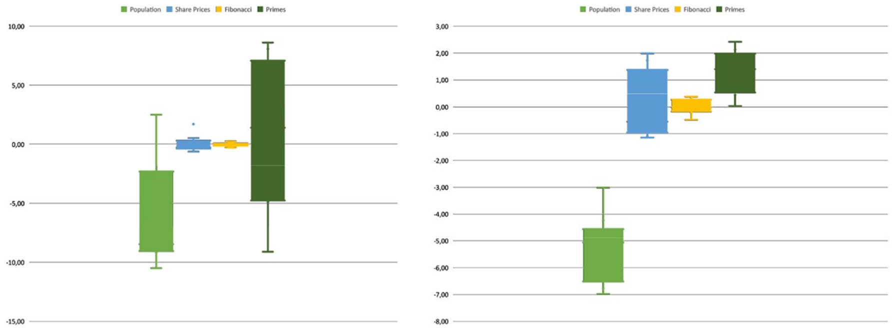

Fig. 1 Plots summarizing the values for the statistics $R_{1,j}$ and $R_{2,j}$, respectively

Table 6 p-Values for the tests $IS_{1,E}$, $IS_{1,M}$, $IS_{2,E}$, and $IS_{2,M}$ applied on our illustrative data sets

	Data sets			
	Fibonacci	Primes	Population	Share prices
Test \ n	1000	1000	3998	369
$IS_{1,E}$	1.00	0.00	0.00	0.89
$IS_{1,M}$	1.00	0.00	0.00	0.88
$IS_{2,E}$	1.00	0.02	0.00	0.26
$IS_{2,M}$	1.00	0.00	0.00	0.33

3 Illustration

We illustrate our methods by four carefully selected data sets. The first two data sets are chosen to illustrate that our tests really yield results conforming to Number Theory. The other two represent empirical data sets.

#1: Fibonacci ($n = 1000$) The Fibonacci numbers are proved to be Benford distributed, cf. e.g. Berger and Hill (2015).

#2: Prime Numbers ($n = 1000$) Opposite to the Fibonacci numbers the prime numbers are known to be not Benford, cf. e.g. Berger and Hill (2015).

#3: Population ($n = 3998$) This data set consists of the number of inhabitants in cities worldwide that are larger than 100.000 people. It illustrates that data from certain truncated distributions are not Benford, cf. Appendix D.

#4: Share Prices ($n = 369$) The data include share prices as a mixture from international stock market indices. Such data sets are assumed to behave like Benford, according to the Theorem of Mixtures due to Berger and Hill (2015, section 8.3).

First, we study the behaviour of each of the four invariant sum tests. The level of significance is $\alpha = 0.01$. The nine values of the statistics $R_{1,i}$, $i = 1, \ldots, 9$ as well as the ten values of the statistics $R_{2,i}$, $i = 0, \ldots, 9$ are summarized in Fig. 1. We see that the values of $R_{1,j}$ for the Share Prices and for the Fibonacci numbers are very close to zero which provides some evidence of the Benford property. For the datasets Population and Prime Numbers the boxes are large and far from zero which gives some evidence of non-Benford. For the second significant digit it is similar but sometimes less clear. However, for Share Prices most of the values $R_{2,j}$ are less than one resulting in small values for $IS_{2,E}$ and $IS_{2,M}$. Table 6 contains the p-values for the tests $IS_{1,E}$, $IS_{1,M}$, $IS_{2,E}$, and $IS_{2,M}$.

Note that the values are rounded. This way, the entries especially for p-values may become 1.00 or 0.00. The p-value of (nearly) 1.00 of Fibonacci numbers signals evidence of the from Number Theory well-known fact that they are nearly perfect Benford. The (rounded) p-value of 0.00 gives very strong evidence of the well known fact that prime numbers are not Benford, also known from number theory. For the notation of evidence and (very) strong evidence we refer to Wasserman (2004). The two data sets, Fibonacci and Prime numbers, are selected for illustrating that all the tests considered yield a decision that confirms the mathematical theory. Note that for the Fibonacci and prime numbers we have some few entries with only one digit. As they represent structural non-existing items they are removed when testing for the second significant digit.

The tests conform to the underlying theories, i.e number theory, Berger and Hill's theorem on mixtures and the conjecture of bounded domains in Appendix D. The data set #1 (Fibonacci) is clearly Benford. Furthermore, data set #3 (Population) is clearly not Benford. For an explanation based on trimming of values or bounded domains we refer to the Appendix D. Prime numbers (data set #2) are known to be not Benford which is clearly illustrated by the three tests $IS_{1,E}, IS_{1,M}$ and $IS_{2,M}$, the test $IS_{2,E}$ does not reject Benford at the $\alpha = 0.01$ level that might be caused by less power of $IS_{2,E}$ for sample size n=1000. The data set

Table 7 Critical values ($\alpha = 0.01$) and observed values of the various Goodness of Fit tests

Test	Critical	Data sets			
		Population	Share Prices	Fibonacci[1]	Primes[1]
	Value	n=3998	n=369	n=1000	n=1000
KS1	1.42[2]	**15.4**	0.36	0.03	**5.41**
KS2	1.46	**5.00**	0.58	0.20	1.41
GoF1	20.09	**1090**	3.45	0.17	**299.9**
GoF2	21.67	**136**	9.06	0.58	11.22
MAD1	3.60	**30.7**	1.23	0.23	**14.9**
MAD2	3.92	**10.0**	2.60	0.79	2.91
$IS_{1,E}$	22.64	**1274**	4.10	0.18	**332.8**
$IS_{1,M}$	21.67	**1293**	4.51	0.34	**303.2**
$IS_{2,E}$	23.30	**308**	12.1	0.69	21.6
$IS_{2,M}$	23.11	**918**	11.3	0.65	**54.7**

[1] For the tests KS2, GoF2, MAD2, $IS_{2,E}$, and $IS_{2,M}$ we removed all entries with only one digit

[2] The critical value for the KS1-test is obtained by Morrow (2014)

#4 (Share Prices) is in accordance with the Mixture Theorem of Berger and Hill (2015) such leading to Benford's law.

The results for all tests considered, KS1, KS2, GoF1, GoF2, MAD1, MAD2, $IS_{1,E}$, $IS_{1,M}$, $IS_{2,E}$ and $IS_{2,M}$, are presented in Table 7. Bold values mean 'rejection', given $\alpha = 0.01$. Note that for that decision one and only one corresponding test himself is considered, such that the multiple test problem is not relevant here. The classical tests GoF1, KS1, and MAD1 confirm the results of the invariant sum tests.

Note that when testing primes most of the tests based on the second significant digit do not reject Benford due to low power. However, if we inccrease the sample size and take all prime numbers between 11 and 100,000 then Benford will be rejected by all the tests based on the second significant digit, too.

4 Summary

We consider several statistical tests of the Benford law, some few are known, most are new. Completely new tests are that based on the second significant digit, except test GoF2. The various variants of the invariant sum tests are appealing as they use the significand. Therefore the Invariant Sum tests use the full information of the data.

We have shown that almost all the tests give confirmative results for data sets for which there is a theory whether the Benford property is true or not, except for primes with the second significant digit, cf. Tables 7 and 8. The last line in Table 8 presents the Bonferroni adjusted p-values and it is intended only for a very quick impression to the reader. We see that data sets #1 and, quite sure, #4 are Benford, the other two are not.

Table 8 p-Values of the various test statistics

Test	Data sets			
	Population	Share prices	Fibonacci[a]	Primes[a]
	n = 3998	n = 369	n = 1000	n = 1000
KS1	0.000	0.890	1.000	0.000
KS2	0.000	0.608	0.968	0.013
GoF1	0.000	0.903	1.000	0.000
GoF2	0.000	0.432	1.000	0.261
MAD1	0.000	0.946	1.000	0.000
MAD2	0.000	0.351	1.000	0.192
$IS_{1,E}$	0.000	0.888	1.000	0.000
$IS_{1,M}$	0.000	0.875	1.000	0.000
$IS_{2,E}$	0.000	0.261	1.000	0.017
$IS_{2,M}$	0.000	0.334	1.000	0.000
BON	0.000	1.000	1.000	0.000

[a] For the tests KS2, GoF2, MAD2, $IS_{2,E}$, and $IS_{2,M}$ we removed all entries with only one digit

Table 9 Variances of $S(X)\mathbb{1}_{C(d_1)}(S(X))$ if X is Benford

First digit d_1								
1	2	3	4	5	6	7	8	9
0.463	0.897	1.331	1.766	2.200	2.634	3.069	3.503	3.937

Table 10 Correlation matrix $\boldsymbol{\Sigma}_{1,R}$ of $\mathrm{Sum}_{1,j}$ of the Invariant sum tests $IS_{1,E}$ and $IS_{1,M}$ if X is Benford

1.	−0.293	−0.240	−0.209	−0.187	−0.171	−0.158	−0.148	−0.140
−0.293	1.	−0.173	−0.150	−0.134	−0.123	−0.114	−0.106	−0.100
−0.240	−0.173	1.	−0.123	−0.110	−0.101	−0.093	−0.087	−0.082
−0.209	−0.150	−0.123	1.	−0.096	−0.087	−0.081	−0.076	−0.072
−0.187	−0.134	−0.110	−0.096	1.	−0.078	−0.073	−0.068	−0.064
−0.171	−0.123	−0.101	−0.087	−0.078	1.	−0.066	−0.062	−0.059
−0.158	−0.114	−0.093	−0.081	−0.073	−0.066	1.	−0.058	−0.054
−0.148	−0.106	−0.087	−0.076	−0.068	−0.062	−0.058	1.	−0.051
−0.140	−0.100	−0.082	−0.072	−0.064	−0.059	−0.054	−0.051	1.

In future research it is intended to investigate which of the considered tests is good for various alternative directions. Moreover, various sample sizes are to be considered. Furthermore, we intend to construct tests that are based on sum invariance and on other invarianvce principles.

Appendix A

Expectations, variances and covariances of the significands with fixed first or fixed second digit, respectively

If the random variable X is Benford, then it has sum invariant significant digits. Let $d_1 \in \{1, 2, ..., 9\}$ be given, then we have

$$\mathbb{E}\big(S(X)\mathbb{1}_{C(d_1)}(S(X))\big) = \int_{d_1}^{d_1+1} t \cdot \frac{1}{t \ln 10} dt = \frac{1}{\ln 10}$$

$$\mathbb{E}(S(X)\mathbb{1}_{C(d_1)}(S(X)))^2 = \int_{d_1}^{d_1+1} t^2 \cdot \frac{1}{t \ln 10} dt = \frac{2d_1 + 1}{2 \ln 10}$$

$$\operatorname{var}\big(S(X)\mathbb{1}_{C(d_1)}(S(X))\big) = \frac{2d_1 + 1}{2 \ln 10} - \Big(\frac{1}{\ln 10}\Big)^2$$

$$\operatorname{cov}\big(S(X)\mathbb{1}_{C(d_1)}(S(X)), S(X)\mathbb{1}_{C(d'_1)}(S(X))\big) = -\frac{1}{(\ln 10)^2} \quad \text{if} \quad d_1 \neq d'_1$$

Table 11 Variances of $S(X)\mathbb{1}_{C_2(d_2)}(S(X))$ if X is Benford

second digit d_2									
0	1	2	3	4	5	6	7	8	9
1.821	1.860	1.899	1.938	1.977	2.017	2.056	2.095	2.134	2.173

which are already well-known results, cf. Barabesi et al. (2021).

The variances of $S(X)\mathbb{1}_{C(d_1)}(S(X))$ are tabulated in Table 9 and the correlation matrix $\mathbf{\Sigma}_{1,R}$ of the vector $\mathbf{R}_1$ is tabulated in Table 10.

Now, let $d_2 \in \{0, 1, 2, ..., 9\}$ be given. Recall the sets $C_2(d_2) = \{x \in [1, 10) : D_2(x) = d_2\}$. Then we have

$$\begin{aligned}
\mathbb{E}\big(S(X)\mathbb{1}_{C_2(d_2)}(S(X))\big) &= \sum_{d_1=1}^{9} \int_{d_1+\frac{d_2}{10}}^{d_1+\frac{d_2+1}{10}} t \cdot \frac{1}{t \ln 10} dt \\
&= \frac{1}{\ln 10} \sum_{d_1=1}^{9} \left(d_1 + \frac{d_2+1}{10} - \left(d_1 + \frac{d_2}{10} \right) \right) = \frac{9}{10 \ln 10} \\
\mathbb{E}\big(S(X)\mathbb{1}_{C_2(d_2)}(S(X))\big)^2 &= \sum_{d_1=1}^{9} \int_{d_1+\frac{d_2}{10}}^{d_1+\frac{d_2+1}{10}} t^2 \cdot \frac{1}{t \ln 10} dt \\
&= \frac{1}{2 \ln 10} \sum_{d_1=1}^{9} \left(\left(d_1 + \frac{d_2+1}{10} \right)^2 - (d_1 + \frac{d_2}{10})^2 \right) \\
&= \frac{1}{2 \ln 10} \sum_{d_1=1}^{9} \left(\frac{2}{10}(d_1 + \frac{d_2}{10} \right) + \frac{1}{10^2} \\
&= \frac{1}{200 \ln 10} \sum_{d_1=1}^{9} \big(20 \cdot d_1 + 2d_2 + 1 \big) = \frac{9(101 + 2d_2)}{200 \ln 10} \\
\mathrm{var}\,\big(S(X)\mathbb{1}_{C_2(d_2)}(S(X))\big) &= \frac{9(101 + 2d_2)}{200 \ln 10} - \left(\frac{9}{10 \ln 10} \right)^2
\end{aligned}$$

The variances of $S(X)\mathbb{1}_{C_2(d_2)}(S(X))$ are tabulated in Table 11.

To obtain the covariance note that the sets $C_2(d_2)$ and $C_2(d_2')$ are disjunct if $d_2 \neq d_2'$, and therefore

$$\mathbb{E}\Big((S(X)\mathbb{1}_{C_2(d_2)}(S(X)) \cdot (S(X)\mathbb{1}_{C_2(d_2')}(S(X)) \Big) = 0 \quad \text{if} \quad d_2 \neq d_2'$$

Therefore the covariance equals

$$\mathrm{cov}\,\big(S(X)\mathbb{1}_{C_2(d_2)}(S(X)), S(X)\mathbb{1}_{C_2(d_2')}(S(X)) \big) = -\frac{81}{(10 \ln 10)^2} \quad \text{if} \quad d_2 \neq d_2'.$$

Therefore the correlation matrix $\mathbf{\Sigma}_{R,2}$ of the vector $\mathbf{R}_2$ can be computed, see Table 12.

Table 12 Correlation matrix $\mathbf{\Sigma}_{2,R}$ of $\mathrm{Sum}_{2,j}$ of the Invariant sum tests $IS_{2,E}$ and $IS_{2,M}$ if X is Benford

1.	−0.081	−0.080	−0.080	−0.079	−0.078	−0.077	−0.077	−0.076	−0.075
−0.081	1.	−0.080	−0.079	−0.078	−0.077	−0.077	−0.076	−0.075	−0.075
−0.080	−0.080	1.	−0.078	−0.077	−0.077	−0.076	−0.075	−0.074	−0.074
−0.080	−0.079	−0.078	1.	−0.077	−0.076	−0.075	−0.074	−0.074	−0.073
−0.079	−0.078	−0.077	−0.077	1.	−0.075	−0.074	−0.074	−0.073	−0.072
−0.078	−0.077	−0.077	−0.076	−0.075	1.	−0.074	−0.073	−0.073	−0.072
−0.077	−0.077	−0.076	−0.075	−0.074	−0.074	1.	−0.072	−0.072	−0.071
−0.077	−0.076	−0.075	−0.074	−0.074	−0.073	−0.072	1.	−0.071	−0.070
−0.076	−0.075	−0.074	−0.074	−0.073	−0.072	−0.072	−0.071	1.	−0.070
−0.075	−0.075	−0.074	−0.073	−0.072	−0.072	−0.071	−0.070	−0.070	1.

Appendix B

Proof of theorems 2 and 3

Proof of theorem 2 To obtain the asymptotic distributions of $IS_{1,E}$ and $IS_{1,M}$ let $\mathbf{U}_1$ be the matrix of Eigenvectors of the asymptotic correlation matrix $\mathbf{\Sigma}_{1,R}$. Let $\mathbf{\Lambda}_1 = \operatorname{diag}(\lambda_{1,1}, \ldots, \lambda_{1,9})$, where the $\lambda_{1,j}$, $j = 1, \ldots, 9$, are the Eigenvalues of $\mathbf{\Sigma}_{1,R}$.

Consider the random vector

$$\mathbf{W}_1^* = \mathbf{\Lambda}_1^{-\frac{1}{2}} \mathbf{U}_1' \mathbf{R}_1.$$

Obviously, $\operatorname{cov}(\mathbf{W}_1^*) = \mathbf{\Lambda}_1^{-\frac{1}{2}} \mathbf{U}_1' \mathbf{\Sigma}_{1,R} \mathbf{U}_1 \mathbf{\Lambda}_1^{-\frac{1}{2}} = \mathbf{I}_1$, where $\mathbf{I}_1$ is the (9×9) identity matrix. Let $\mathbf{0}_1$ be the null vector of dimension 9. Therefore $\mathbf{W}_1^* \sim \mathcal{N}(\mathbf{0}_1, \mathbf{I}_1)$, asymptotically, under $H_{1,0}$. This way we have

$$\begin{aligned} IS_{1,E} &= \mathbf{R}_1'\mathbf{R}_1 = \mathbf{R}_1'\mathbf{U}_1\mathbf{\Lambda}_1^{-\frac{1}{2}}\mathbf{\Lambda}_1\mathbf{\Lambda}_1^{-\frac{1}{2}}\mathbf{U}_1'\mathbf{R}_1 = \mathbf{W}_1^{*\prime}\mathbf{\Lambda}_1\mathbf{W}_1^* = \sum_{j=1}^{9} \lambda_{1,j} W_{1,j}^2 \\ IS_{1,M} &= \mathbf{R}_1'\mathbf{\Sigma}_{1,R}^{-1}\mathbf{R}_1 = \mathbf{R}_1'\mathbf{U}_1\mathbf{\Lambda}_1^{-1}\mathbf{U}_1'\mathbf{R}_1 = \mathbf{R}_1'\mathbf{U}_1\mathbf{\Lambda}_{1,R}^{-1/2}\mathbf{\Lambda}_1^{-1/2}\mathbf{U}_1'\mathbf{R}_1 = \mathbf{W}_1^{*\prime}\mathbf{W}_1^* \\ &= \sum_{j=1}^{9} W_{1,j}^2 \end{aligned}$$

where the $W_{1,j}$ are the components of the vectors $\mathbf{W}_1^*$. Therefore the statistics $IS_{1,E}$ are, under $H_{0,1}$, asymptotically weighted sums of independent χ_1^2 distributed random variables, where the weights $\lambda_{1,j}$ are the Eigenvalues of $\mathbf{\Sigma}_{1,R}$. Since the statistics $IS_{1,M}$ are asymptotically sums of nine squares of independent standard normal random variables, we have $IS_{1,M} \sim \chi_9^2$. □

Table 13 Eigenvalues of the correlation matrix $\mathbf{\Sigma}_{1,R}$ in the Invariant Sum Tests $IS_{1,E}$ ans $IS_{1,M}$

1.329	1.181	1.125	1.096	1.078	1.066	1.057	1.050	0.019

Table 14 Eigenvalues of the correlation matrix $\mathbf{\Sigma}_{2,R}$ in the Invariant Sum Tests $IS_{2,E}$ ans $IS_{2,M}$

1.082	1.080	1.078	1.077	1.075	1.074	1.072	1.071	1.069	0.323

Proof of theorem 3 To obtain the asymptotic distributions of $IS_{2,E}$ and $IS_{2,M}$ let $\mathbf{U}_2$ be the matrix of Eigenvectors of the asymptotic correlation matrix $\mathbf{\Sigma}_{2,R}$. Let $\mathbf{\Lambda}_2 =$ diag $(\lambda_{2,0}, \dots, \lambda_{2,9})$, where the $\lambda_{2,j}$, $j = 0, \dots, 9$, are the Eigenvalues of $\mathbf{\Sigma}_{2,R}$.

Consider the random vector

$$\mathbf{W}_2^* = \mathbf{\Lambda}_2^{-\frac{1}{2}} \mathbf{U}_2' \mathbf{R}_2.$$

Obviously, cov $(\mathbf{W}_2^*) = \mathbf{\Lambda}_2^{-\frac{1}{2}} \mathbf{U}_2' \mathbf{\Sigma}_{2,R} \mathbf{U}_2 \mathbf{\Lambda}_2^{-\frac{1}{2}} = \mathbf{I}_2$, where $\mathbf{I}_2$ is the (10×10) identity matrix. Let $\mathbf{0}_2$ be the null vector of dimension 10. Therefore $\mathbf{W}_2^* \sim \mathcal{N}(\mathbf{0}_2, \mathbf{I}_2)$, asymptotically, under $H_{2,0}$. This way we have

$$\begin{aligned} IS_{2,E} =& \mathbf{R}_2'\mathbf{R}_2 = \mathbf{R}_2'\mathbf{U}_2\mathbf{\Lambda}_2^{-\frac{1}{2}}\mathbf{\Lambda}_2\mathbf{\Lambda}_2^{-\frac{1}{2}}\mathbf{U}_2'\mathbf{R}_2 = \mathbf{W}_2^{*\prime}\mathbf{\Lambda}_2\mathbf{W}_2^* = \sum_{j=0}^{9} \lambda_{2,j} W_{2,j}^2 \\ IS_{2,M} =& \mathbf{R}_2'\mathbf{\Sigma}_{2,R}^{-1}\mathbf{R}_2 = \mathbf{R}_2'\mathbf{U}_2\mathbf{\Lambda}_2^{-1}\mathbf{U}_2'\mathbf{R}_2 = \mathbf{R}_2'\mathbf{U}_2\mathbf{\Lambda}_{2,R}^{-1/2}\mathbf{\Lambda}_2^{-1/2}\mathbf{U}_2'\mathbf{R}_2 = \mathbf{W}_2^{*\prime}\mathbf{W}_2^* \\ =& \sum_{j=0}^{9} W_{2,j}^2 \end{aligned}$$

where the $W_{2,j}$ are the components of the vectors $\mathbf{W}_2^*$. Therefore the statistics $IS_{2,E}$ are, under $H_{0,2}$ asymptotically weighted sums of independent χ_1^2 distributed random variables, where the weights $\lambda_{2,j}$ are the Eigenvalues of $\mathbf{\Sigma}_{2,R}$. Since the statistics $IS_{2,M}$ are asymptotically sums of ten squares of independent standard normal random variables, we have $IS_{2,M} \sim \chi_{10}^2$. □

Appendix C

Approximation of the weighted sums by a χ^2 distributed random variable

The quadratic forms $\mathbf{R}_k'\mathbf{R}_k$, $k = 1, 2$, will be approximated by (possibly noncentral) χ^2 distributed random variables Z_k suitably shifted and scaled according to the idea of Liu et al. (2009). It is based on the moment equating method. The degrees of freedom, the location and scale parameters and the noncentrality parameter are to be determined. Recall that $\lambda_{1,j}, j = 1, \dots, 9$ are the Eigenvalues of the correlation matrix $\mathbf{\Sigma}_{1,R}$. The

Eigenvalues can be found in Table 13. Analogously, recall that $\lambda_{2,j}, j = 0, \ldots, 9$ are the Eigenvalues of the correlation matrix $\mathbf{\Sigma}_{2,R}$. The Eigenvalues can be found in Table 14.

Denote

$$c_{1,r} = \sum_{j=1}^{9} \lambda_{1,j}^{r}, \qquad c_{2,r} = \sum_{j=0}^{9} \lambda_{2,j}^{r}, \qquad r = 1, 2, 3, 4.$$

Consider first the case of the first significant digit (k = 1), and denote

$$s_{1,1} = \frac{c_{1,3}}{c_{1,2}^{3/2}} = 0.357 \qquad \text{and} \qquad s_{1,2} = \frac{c_{1,4}}{c_{1,2}^{2}} = 0.128.$$

The approximation generally depends on whether we have $s_{1,1}^2 < s_{1,2}$ or not. In our case $s_{1,1}^2 < s_{1,2}$ is true and applying the approximation of Liu et al. (2009) we obtain that the noncentrality parameter of the χ^2 approximation is zero, and the degrees of freedom df_1, and the regression coefficients $\beta_{1,0}$ and $\beta_{1,1}$ are

$$\begin{aligned} df_1 &= \frac{1}{s_{1,1}^2} = 7.84619 \\ \beta_{1,0} &= -\frac{c_{1,2}^2}{c_{1,3}} + c_{1,1} = 0.0780258 \\ \beta_{1,1} &= \frac{c_{1,3}}{c_{1,2}} = 1.13711 \end{aligned}$$

The approximation of our statistic $IS_{1,E}$ is then

$$IS_{1,E} = \mathbf{R}_1'\mathbf{R}_1 \approx \beta_{1,1} Z_1 + \beta_{1,0}, \qquad \text{where} \quad Z_1 \sim \chi^2_{df_1}$$

Therefore, if we choose $\alpha = 0.01$, the critical value of the test $IS_{1,E}$ is $d_{crit,IS_{1,E}} = \beta_{1,1}\chi^2_{1-\alpha,df_1} + \beta_{1,0} = 22.6435$, which is close to the critical value $d_{crit,IS_{1,M}} = \chi^2_{0.99,9} = 21.666$ of the $IS_{1,M}$-test.

In the case of the second significant digit (k = 2) denote

$$s_{2,1} = \frac{c_{2,3}}{c_{2,2}^{3/2}} = 0.3334 \qquad \text{and} \qquad s_{2,2} = \frac{c_{2,4}}{c_{2,2}^{2}} = 0.111117.$$

Again, we have the simpler case, now $s_{2,1}^2 < s_{2,2}$, and the χ^2 approximation is computed in the same way as above,

$$
\begin{aligned}
df_2 =& \frac{1}{s_{2,1}^2} = 8.99964 \approx 9 \\
\beta_{2,0} =& -\frac{c_{2,2}^2}{c_{2,3}} + c_{2,1} = 0.000128938 \approx 0 \\
\beta_{2,1} =& \frac{c_{2,3}}{c_{2,2}} = 1.07527
\end{aligned}
$$

The approximation of our statistic $IS_{2,E}$ is then

$$IS_{2,E} = \mathbf{R}_2'\mathbf{R}_2 \approx \beta_{2,1}Z_2 + \beta_{2,0}, \qquad \text{where} \quad Z_2 \sim \chi^2_{df_2}$$

Therefore, if we choose $\alpha = 0.01$, the critical value of the test $IS_{2,E}$ is approximately $d_{crit,IS_{2,E}} = \beta_{2,1}\chi^2_{1-\alpha,df_2} + \beta_{2,0} = 23.2963$, which is very close to the critical value $d_{crit,IS_{2,M}} = \chi^2_{0.99,10} = 23.2093$ of the $IS_{2,M}$-test.

Appendix D

On the (non-existing) Benford property for conditional distributions conditioned under $X > t$ with large t and with small probability mass $P(X > t)$. This section is intended to illustrate that data set Population is not Benford. Recall that only cities with more than 100,000 inhabitants are considered. Moreover, there are much less cities with more than 100,000 inhabitants than that with less inhabitants. Therefore, for the random variable, say X, with support $[a, \infty)$ we have an underlying conditional distribution, conditioned under $X > 100,000$. Note that the starting point a of the distribution is small in our example we have, perhaps $a = 1$ (inhabitant) or $a = 10$ or $a = 100$), $a \ll 100,000$.

Now, let t be a large threshold ($t = 100,000$ in our example), and let $F(x)$ be a continuous cdf with positive density on support $[a, \infty)$ where $a \ll t$ is some positive real number much less than t and most of the probability mass of the random variable X is below the threshold t. Then the conditional cdf $F(x \mid t) = P(X < x \mid X > t) = \frac{F(x)-F(t)}{1-F(t)}$ may be approximated by the cdf of a Generalized Pareto Distribution (GPD), as a result of the Pickands-Balkema-de Haan theorem, cf. Pickands (1975, theorem 7) or Balkema and Haan (1974). Let $x = t + y, y \geq 0$.

The cdf of the GPD $G(y;k,\sigma)$ is given by

$$GPD(y;k,\sigma) = \begin{cases} 1 - e^{-\frac{y}{\sigma}} & \text{if} \quad k = 0 \\ 1 - \left(1 - \frac{ky}{\sigma}\right)^{\frac{1}{k}} & \text{if} \quad k \neq 0, \end{cases}$$

where k is the shape parameter and σ is the scale parameter. The range of the GPD is given by $0 < y < \infty$ if $k \leq 0$, and $0 < y < \frac{\sigma}{k}$ if $k > 0$, cf. e.g. Smith (1987, p.1175). The parameters k and σ are given by the extreme value theory, cf. e.g. Falk (1989) or Kössler (1999). Note that the parameter $-k$ is sometimes called the extreme value index of the underlying distribution, cf. Haan and Fereira (2006).

Since the cdf F has support $[a, \infty)$ and we consider $t \gg a > 0$ we only have one of the cases $k \leq 0$. We assume a polynomial decreasing density for $x \to \infty$. That is why we may assume that the parameter $k < 0$. Then we have $\sigma = -kt$, cf. e.g. Falk (1989) or Kössler (1999). The conditional cdf $F(x \mid t) = F(t + y \mid t) = F_t(y)$ is approximated by

$$F_t(y) = F_t(x - t) \approx 1 - \left(1 + \frac{y}{t}\right)^{\frac{1}{k}} = 1 - \left(\frac{x}{t}\right)^{\frac{1}{k}}, \qquad (x > t, y > 0)$$

which is a Pareto cdf with scale parameter t and shape parameter $\gamma := \frac{1}{k}$. To obtain the probability $P(D_1(X) = 1) \mid X > t)$ that the first significand has value one, let, for simplicty and without loss of generality, be $t = 10^m$ (in our example we have $m = 5$). Let $G(x) := 1 - \left(\frac{x}{t}\right)^{\gamma}$. We have

$$\begin{aligned} P(D_1(X) = 1 \mid X > t) \approx & \sum_{j=-\infty}^{\infty} \left(G(2 \cdot 10^j) - G(10^j)\right) \\ = & \sum_{j=m}^{\infty} \left((10^{j-m})^{\gamma} - (2 \cdot 10^{j-m})^{\gamma}\right) = \sum_{j=0}^{\infty} \left((10^j)^{\gamma} - (2 \cdot 10^j)^{\gamma}\right) \\ = & (1 - 2^{\gamma}) \sum_{j=0}^{\infty} (10^{\gamma})^j = \frac{1 - 2^{\gamma}}{1 - 10^{\gamma}} := g(\gamma). \end{aligned}$$

Looking at the shape of the function $g(\gamma), \gamma < 0$ we see that for small values of γ the probability $P(D_1(X) = 1 \mid X > t)$ is much larger than the Benford probability of approximately 0.301. For example, for the Pareto with shape parameter $\gamma = -1$ or for the Cauchy distribution we have $k = \gamma = -1$ and the last probability becomes $\frac{5}{9} \approx 0.55$. For the shorter tail Pareto with shape parameter $\gamma = -2$ $(k = -0.5)$ we have $g(\gamma) = g(-2) \approx 0.75$. Even for the very long-tail Pareto with $k = -2$ we obtain $P(D_1(X) = 1) \approx 0.428$ which is still very far from the Benford probability of approximately 0.301.

Note that in the case of an exponential distribution, which is an example for the case of shape parameter $k = 0$, a similar computation yields values for $P(D_1(X) = 1 \mid X > t) \gg 0.301$.

Consider the second significant digit. A similar but somewhat more laborious computation shows that

Table 15 Probabilities $P(D_2(X) = l \mid X > t), l = 0, \ldots, 9$ for various values of γ of the Pareto distribution

γ	k	l									
		0	1	2	3	4	5	6	7	8	9
– 2	−0.5	0.213	0.167	0.133	0.109	0.090	0.076	0.065	0.056	0.049	0.043
-1	– 1	0.156	0.138	0.122	0.109	0.098	0.089	0.081	0.074	0.068	0.063
−0.5	– 2	0.137	0.125	0.115	0.107	0.099	0.093	0.088	0.083	0.079	0.075
−0.25	– 4	0.128	0.119	0.112	0.106	0.100	0.095	0.091	0.087	0.083	0.080

$$
\begin{aligned}
P(D_2(X) = l \mid X > t) &\approx \sum_{j=-\infty}^{\infty} \sum_{n=1}^{9} \big(G((n+1) \cdot (10^j + l)) - G(n \cdot (10^j + l)) \big) \\
&= \sum_{j=-1}^{\infty} \sum_{n=1}^{9} \Big((10^j (10n + l))^\gamma - (10^j (10n + l + 1))^\gamma \Big) \\
&= \sum_{n=1}^{9} \big((10n + l)^\gamma - (10n + l + 1)^\gamma \big) \sum_{j=-1}^{\infty} (10^\gamma)^j \\
&= \sum_{n=1}^{9} \big((10n + l)^\gamma - (10n + l + 1)^\gamma \big) \frac{1}{10^\gamma \cdot (1 - 10^\gamma)}
\end{aligned}
$$

In Table 15 the probabilities $P(D_2(X) = l \mid X > t), l = 0, \dots, 9$ are presented for various values of the parameter γ of the Pareto distribution. It seems that, if the tails of the density are very long as it is the case for small values of k, the distribution of the second significant digit may be closer to Benford.

Appendix E

Frequencies of first and second significant digits

For the convenience of the reader who is interested in reproducing also the classical goodnes-of-fit tests we present the frequencies of the first and second significant digits, rspectively. The frequencies of the second significant digit are obtained after removing all entries with only one digit. Additionally, we present the values for the GoF1 and GoF2 statistics, respectively (Tables 16, 17).

Table 16 Sample sizes n for first digit, frequencies of the first significant digit, and values of the GoF1 statistic

Data set		n	First significant digit									GoF1
			1	2	3	4	5	6	7	8	9	
#1	Fibonacci	1000	301	177	125	96	80	67	56	53	45	0.17
#2	Primes	1000	160	146	139	139	131	135	118	17	15	299.9
#3	Population	3998	2103	775	352	247	165	134	77	77	68	1090
#4	Share Prices	369	107	63	47	34	38	23	20	21	16	3.45

Table 17 Sample sizes n for second digit, frequencies of the second significant digit, and values of the GoF2 statistic

Data set		n	Second significant digit										GoF2
			0	1	2	3	4	5	6	7	8	9	
#1	Fibonacci	994	119	115	103	107	102	95	93	92	86	82	0.58
#2	Primes	996	105	91	104	105	95	104	104	102	94	92	11.22
#3	Population	3998	589	594	483	436	387	374	306	307	256	266	136.0
#4	Share Prices	369	35	45	51	45	30	35	31	38	28	31	9.06

Appendix F

Algorithm that computes the test statistics and the *p*-values of the invariant sum tests

To perform the invariant sum tests, we used the packages SAS, cf. SAS Institute (2022) and Mathematica 12.0, cf. Wolfram Research (2023).

```
Algorithm Invariant Sum tests

input: data x, E(Sum_1), Var(Sum_1), Corr(Sum_1)
               E(Sum_2), Var(Sum_2), Corr(Sum_2)
*      (data, expectations, variances and covariances)
output: IS_1E, IS_2E, IS_1M, IS_2M, p_1E, p_2E, p_1M, p_2M
*      (test statistics and p-values)
usepackages: SAS 9.4, Mathematica 12.0

* Data step and procedures SORT and  MEANS from SAS package
DATA Sums;
Compute Significands S=S(x)
Compute first significant digit i
Compute second significant digit j
RUN;
* Sort according first and second digit, respectively
* SAS requires sorting
PROC SORT data=Sums, out=Sums1; by i; run;
PROC SORT data=Sums, out=Sums2; by j; run;
* Compute the corresponding sums for all i=1 to 9 and for all j=0 to 9

Sum_1i<-PROC MEANS Data=Sums1 SUM; var S; by i;
Sum_2j<-PROC MEANS Data=Sums2 SUM; var S; by j;

Transfer the sums Sum_1i and Sum_2j to mathematica 12.0

Standardize all the sums and obtain the R_1i and R_2j
* compute the test statistics
IS_1E<-Transpose[vec(R_1i)].vec(R_1i)
IS_1M<-Transpose[vec(R_1i)].Inverse[Corr(Sum1)].vec(R_1i)
IS_2E<-Transpose[vec(R_2j)].vec(R_2j)
IS_2M<-Transpose[vec(R_2j)].Inverse[Corr(Sum2)].vec(R_2j)
* p-values step, Mahalanobis version
p_1M<-Quantile[ChiSquareDistribution[9],IS_1M]
p_2M<-Quantile[ChiSquareDistribution[10],IS_2M]
*p-values step Euclidean version
Compute approximations, degrees of freedom df1, df2, regression coefficients
beta_10, beta_11, beta_20, beta_21, cf. Appendix C
p_1E<-Quantile[ChiSquareDistribution[df1],(IS_1E-beta_10)/beta_11]
p_2E<-Quantile[ChiSquareDistribution[df2],(IS_2E-beta_20)/beta_21]
end Invariant Sum Tests
```

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflicts of interest There are no conflicts of interest.

References

Allaart PC (1997) An invariant sum characterization of Benford's law. J Appl Prob 34(1):288–291

Balkema A, Haan L (1974) Residual life time at great age. Ann Probab 2:792–804

Barabesi L, Cerasa A, Cerioli A, Perrotta D (2021) On characterizations and tests of Benford's law. J Am Stat Assoc 117:1887–1903

Benford F (1938) The law of anomalous numbers. Proc Am Philos Soc 78(4):551–572

Berger A, Hill TP (2011) A basic theory of Benford's law. Probab Surv 8:1–126

Berger A, Hill TP (2015) An introduction to Benford's Law. Princeton University Press, Princeton and Oxford

Berger A, Hill TP, Rogers E (2024) Benford Online Bibliography. https://www.benfordonline.net. "[Online; Accessed]"

Cerqueti R, Lupi C (2021) Some new tests of conformity with Benford's law. Stats 4:745–761

Cerqueti R, Maggi M (2021) Data validity and statistical conformity with Benford's law. Chaos Solutions Fractals 144:110740

Darling DA (1957) The Kolmogorov, Cramer-von-Mises test. Ann Math Stat 28(4):823–838

Der Tagesspiegel: So war der Tag: DAX verließen die Kräfte, No.24476, 13.3.2021, p 18

Diekmann A (2007) Not the first digit! using Benford's law to detect fraudulent scientific data. J Appl Stat 34:321–329

Falk M (1989) Best attainable rate of joint convergence of extremes. In: Extreme Value Theory, Hüsler J, Reiss RD, (Ed.) Proceedings of a conference held in Oberwolfach, Dec 6–12, 1987, pp 1–9. Springer, New York

Haan L, Fereira A (2006) Extreme value theory. Springer, New York

Hein J, Zobrist R, Konrad C, Schüpfer G (2012) Scientific fraud in 20 falsified anesthesia papers. Aneasthesist 61(61):543–549

Kazemitabar J, Kazemitabar J (2022) Benford test based on logarithmic property. Int J Audit Technol 4:279–291

Kolmogorov AN (1933) Sulla determinazione empirica di una legge di distribuzione. Giorn dell'Inst Ital degli An 4:83–91

Kössler W (1999) A new one-sided variable inspection plan for continuous distribution functions. Allgemeines Statistisches Archiv 83:416–433

Kössler W, Lenz H-J, Wang XD (2024) Detecting manipulated data sets using Benford's law. In: Knoth S, Schmid W (Eds) Frontiers in statistical quality control, **14**

Liu H, Tang Y, Zhang HH (2009) A new chi-square approximation to the distribution of nonnegative definite quadratic forms in noncentral normal variables. Comput Stat Data Anal 53:853–856

Mebane WR (2010) Election Fraud or Strategic Voting? Can Second-digit Tests Tell the Difference? https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.697.3403 &rep=rep1 &type=pdf

Miller LH (1956) Table of percentage points of Kolmogorov statistics. J Am Stat Assoc 51(273):111–121

Morrow J (2014) Benford's law, families of distributions and a test basis. Discussion Paper No 1291, Centre for Economic Performance, LSE, London

Newcomb S (1881) Note on the frequency of use of the different digits in natural numbers. Am J Math 4(1):39–40

Nigrini MJ (1992) The detection of income evasion through an analysis of digital distributions. PhD thesis, University of Cincinnati

Nigrini MJ (2012) Benford's Law: applications for forensic accounting, auditing, and fraud detection. John Wiley & Sons, Hoboken, New Jersey (2012). https://doi.org/10.1002/9781119203094

Pearson K (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it cab be reasonably supposed to have arisen from random sampling. Phil Ma Ser 5(50):157–175

Pickands J (1975) Statistical inference using extreme order statistics. Ann Stat 3:119–135

Pinkham RS (1961) On the distribution of first significant digits. Ann Math Stat 32(4):1223–1230

SAS Institute Inc.: Base SAS®9.3 Procedures Guide. Cary, NY (2022)

Smirnov NV (1948) Tables for estimating goodness of fit of empirical distribution. Ann Math Stat 19(2):279–281

Smith RL (1987) Estimating tails of probability distributions. Ann Stat 15:1174–1207

UNStats Report https://unstats.un.org/unsd/demographic-social/products/dyb/documents/dyb2016/table08.pdf. "[Online; accessed]" (2016)

Wasserman L (2004) All of Statistics, a concise course in statistical inference. Springer

Wolfram Research, Inc.: Mathematica, Version 12.0. Champaign, IL, 2023. https://www.wolfram.com/mathematica

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.