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

Wolfgang Kössler ( $\checkmark$ koessler@informatik.hu-berlin.de)
Humboldt-Universität zu Berlin
Hans-J. Lenz
Freie Universität Berlin
Xing David Wang
Humboldt-Universität zu Berlin

## Research Article

Keywords: Benford Law, Goodness of Fit Test, Sum Invariance, Data Fraud, Data Manipulation, Aggregation of $p$-values, Data Quality

Posted Date: September 11th, 2023
DOI: https://doi.org/10.21203/rs.3.rs-3336839/v1
License: © (i) This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

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

Wolfgang Kössler ${ }^{1^{*}}$, Hans-J. Lenz ${ }^{2}$ and Xing D. Wang ${ }^{1}$<br>${ }^{1 *}$ Institut für Informatik, Humboldt Universität zu Berlin, Rudower Chaussee 25, Berlin, 12489, Germany, Berlin.<br>${ }^{2}$ Institut für Statistik und Ökonometrie, Freie Universität Berlin, Boltzmannstr. 20, 14195, Germany, Berlin.<br>${ }^{1}$ Institut für Informatik, Humboldt Universität zu Berlin, Rudower Chaussee, Berlin, 12489, Germany, Berlin.

> *Corresponding author(s). E-mail(s): koessler@informatik.hu-berlin.de; Contributing authors: hans-j.lenz@fu-berlin.de; wangxida@informatik.hu-berlin.de;


#### 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 a universe is not uniformly, but logarithmically distributed. Especially, the first non-zero digit $\boldsymbol{D}_{\mathbf{1}}$ is One with probability $\boldsymbol{P}\left(\boldsymbol{D}_{\mathbf{1}}=\mathbf{1}\right)=$ $\log _{10} \mathbf{2} \approx \mathbf{0 . 3}$. There are several tests available for testing Benford, the best known are Pearson's $\boldsymbol{\chi}^{2}$-test, the Kolmogorov-Smirnov test and the MAD-test suggested by Nigrini (2012). The latter test was enhanced to significance tests in Kössler, Lenz and Wang (2021) and in Cerqueti and Lupi (2021). 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 from the sample size. 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; Aggregation of p-values; Data Quality

## 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 $\chi^{2}$ goodness-of-fit test. Not every real or artificial data set follows the Benford Law, the question arises how this can be tested in practice.

Berger and Hill (2011) as well as Nigrini (1992) analyzed the scale-, baseand 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 $2, \ldots, 9$, 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 $\chi^{2}$-distributed.

Emphasis is especially taken on the second significant digit. A $\chi^{2}$ goodness-of-fit test for the second digit was already suggested, cf. eg. Diekmann (2007) [7]. We suggest some further tests based on properties of the second significant digit. In section 2.1 we present some basic properties and some statistical tests for testing Benford that are applied later on. In section 2.2 we recall the $\chi^{2}$ 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 section 2.3 we introduce four variants of the invariant sum test, three of them are new, and in section 3 we illustrate the considered tests on some chosen data sets. In section 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 $d \in\{1,2, \ldots, 9\} \quad$ 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 $S(x)$ of $x \in \mathbb{R}$ is defined as $S(x)=t$ with $t \in[1,10)$ where $|x|=10^{k} t$ if $x \neq 0$, else $S(x):=0$.

Next, we can state when the significand and the first significant digit of a random variable $X$ are distributed according to 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

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

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

In Table 1, we give the distribution of the leading digit $D_{1}$. In the following

Table 1 Probabilities $P\left(D_{1}(X)=d_{1}\right)$ according to Benford's Law

| $d_{1}$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $P\left(d_{1}\right)$ | 0.301 | 0.176 | 0.124 | 0.096 | 0.079 | 0.066 | 0.057 | 0.051 | 0.045 |

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).

### 2.2 Classical tests against Benford and their modifications

Our test problem is in general

$$
H_{0}: X \sim \text { Benford } \quad \text { against } \quad H_{1}: X \nsim \text { Benford. }
$$

Note, $H_{1}$ is a very large class of alternatives.
The $\chi^{2}$-test is one of the most popular goodness-of-fit tests, and it was originated by Pearson (1900). The $\chi^{2}$-test statistic measures the relative distance between the relative frequencies $n_{j} / n$ and the probabilities $p_{j}=P\left(D_{1}=d_{j}\right)$ for all $j=1,2, \ldots, 9$ under the Benford Law, and it is defined by

$$
\begin{equation*}
\chi^{2}=n \sum_{j=1}^{9} \frac{\left(n_{j} / n-p_{j}\right)^{2}}{p_{j}}=\sum_{j=1}^{9} \frac{\left(n_{j}-n p_{j}\right)^{2}}{n p_{j}} \tag{2}
\end{equation*}
$$

The $\chi^{2}$-test rejects the null hypothesis $H_{0}$, if $\chi^{2}>\chi_{1-\alpha, 8}^{2}$, where $\chi_{1-\alpha, 8}^{2}$ is the $(1-\alpha)$ quantile of the $\chi^{2}$ distribution with eight degrees of freedom. Note that the $\chi^{2}$ goodness-of-fit test is an approximate test, the statistic (2) is asymptotically $\chi^{2}$-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 $d \in\{0,1,2, \ldots, 9\}$ is $P\left(D_{2}(X)=d\right)=\sum_{j=1}^{9} \log _{10}\left(1+\frac{1}{10 j+d}\right)$, 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 $\chi^{2}$ goodness-of-fit test by GoF1 and GoF2, respectively.

Table 2 Probabilities $P\left(D_{2}(X)=d_{2}\right)$ according to Benford's Law

| $d_{2}$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $P\left(d_{2}\right)$ | 0.1197 | 0.1139 | 0.1088 | 0.1043 | 0.1003 | 0.0967 | 0.0934 | 0.0904 | 0.0876 | 0.0850 |

An alternative goodness-of-fit test is the Kolmogorov-Smirnov (KS) test, cf. Kolomogorov (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

$$
\begin{equation*}
d_{\max }=\sup _{x \in \mathbb{R}}\left|F_{n}(x)-F_{0}(x)\right| . \tag{3}
\end{equation*}
$$

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). Secondly we apply the test to the second significant digit.

The critical values of the KS test were completely tabulated by Miller (1956)[16] for underlying continuous distributions. Morrow (2014)[17] 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.

Table 3 Asymptotic critical values $c_{K S 2,1-\alpha}$ and $c_{M A D 2,1-\alpha}$ of the KS2 and MAD2 test

|  | $\alpha$ |  |  |
| ---: | ---: | ---: | ---: |
|  | 0.01 | 0.05 | 0.1 |
| KS2 | 1.46 | 1.19 | 1.05 |
| MAD2 | 3.92 | 3.42 | 3.18 |

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

$$
\begin{equation*}
M A D=\sqrt{n} \sum_{j=1}^{k}\left|n_{j} / n-p_{j}\right| \tag{4}
\end{equation*}
$$

which we call MAD-test. It uses a suitably scaled sum of the absolute deviations between the relative frequencies and the Benford probabilities for the first digit. The idea is due to Nigrini (2012), who used a slightly different form, $M A D_{N}=\sum_{j=1}^{k} \frac{\left|n_{j} / n-p_{j}\right|}{k}$, where the index N stays for Nigrini. In its original version it is not a statistical test in its proper sense with acceptance/rejection domains, but maps $M A D_{N}$ to linguistic terms of conformance with Benford's law, cf. Nigrini (2012). Moreover, the $M A D_{N}$-test depends on the current sample size. In our new version we introduced the factor $\sqrt{n}$ to avoid this disadvantage. This property is illustrated in Table 4. The motivation for introducing this factor is that the relative frequencies tend to the true probability with $\sqrt{n}$ rate. Recently, Cerqueti and Lupi (2021) [18] 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 $c_{M A D, 1-\alpha}=3.60$ 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) [7], cf. also Nigrini (2012) [1]. However,

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

|  | $\alpha$ |  |  |
| ---: | :--- | :--- | :--- |
| 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 |
| $\infty$ | 2.869 | 3.084 | 3.485 |

we consider data sets with moderate sample sizes, nearly between $\mathrm{N}=200$ and $\mathrm{n}=4000$. 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.

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) [19] make use of the alternative definition of Benford's Law saying that the logarithms of the significands are uniformly distributed. Cerqueti and Maggi (2021) [20] discuss some distance measures, especially the sum of squares deviation and the MAD.

### 2.3 Invariant Sum tests

### 2.3.1 General Description

In this section we apply the invariant-sum property of Benford, cf. Nigrini (1992), Allaart (1997), and Berger and Hill (2015, theorem 5.18). To do this we define the sets $C\left(d_{1}, \ldots, d_{m}\right)=\left\{x \in[1,10): D_{j}(x)=d_{j}\right.$ for $\left.j=1, \ldots, m\right\}$, $C_{1}\left(d_{1}\right)=\left\{x \in[1,10): D_{1}(x)=d_{1}\right\}=\left[d_{1}, d_{1}+1\right)$ and $C_{2}\left(d_{2}\right)=\{x \in$ $\left.[1,10): D_{2}(x)=d_{2}\right\}=\bigcup_{j=1}^{9}\left[j+\frac{d_{2}}{10}, j+\frac{d_{2}+1}{10}\right) . C\left(d_{1}, \ldots, d_{m}\right)$ is the set of all significands with first $m$ digits $d_{1}, \ldots, d_{m}, C_{1}\left(d_{1}\right)$ is the set of all significands with first digit $d_{1}$, and $C_{2}\left(d_{2}\right)$ 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 $\mathbb{E}\left(S(X) \mathbb{1}_{C\left(d_{1}, \ldots, d_{m}\right)}(S(X))\right)$ are the same for all tuples $\left(d_{1}, \ldots, d_{m}\right), d_{1} \neq 0$ 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 $1,2,3, \ldots, 9$ is the same. The same is true for the expectation of the sum of all significands with second digit $0,1, \ldots, 9$.

Let us start with the first significant digit. Denote by $\theta=$ $\mathbb{E}\left(S(X) \mathbb{1}_{C_{1}(i)}(S(X))\right)=\frac{1}{\ln 10}$ the expectation of the random variable
$S(X) \mathbb{1}_{C_{1}\left(d_{1}\right)}(S(X))$ if Benford is true. Let $\theta_{i}$ 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 $[j, j+1)$

$$
\operatorname{Sum}_{1, j}=\sum_{i=1}^{n} S\left(X_{i}\right) \mathbb{1}_{C_{1}(j)}\left(S\left(X_{i}\right)\right)
$$

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

$$
R_{1, j}=\frac{\operatorname{Sum}_{1, j}-\mathbb{E}\left(\mathrm{Sum}_{1, j}\right)}{\sqrt{\operatorname{var}\left(\mathrm{Sum}_{1, j}\right)}}
$$

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

Let be $\mathbf{R}_{1}=\left(R_{1,1}, \ldots, R_{1,9}\right)$ and $\boldsymbol{\Sigma}_{R_{1}}$ 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

$$
I S_{1, E}=\mathbf{R}_{1}^{\prime} \mathbf{R}_{1} \quad \text { and } \quad I S_{1, M}=\mathbf{R}_{1}^{\prime} \boldsymbol{\Sigma}_{R_{1}}^{-1} \mathbf{R}_{1}
$$

where $I S$ stays for Invariant Sum. The statistic $I S_{1, E}$ is the Euklidean distance of the vector $\mathbf{R}_{1}$ of standardized sums from zero, and $I S_{1, M}$ 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 $I S_{1, M}$ is asymptotically $\chi^{2}$-distributed with nine degrees of freedom, and $I S_{1, E}$ is is approximated by a weighted sum of independent $\chi^{2}$-distributed random variables, each with one degree of freedom.

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 $I S_{1, M}>\chi_{1-\alpha, 9}^{2}$ or if $I S_{1, E}>c_{I S_{1, E}, 1-\alpha}$, respectively, where $\chi_{1-\alpha, 9}^{2}$ is the $1-\alpha$-quantile of the $\chi^{2}$ distribution with nine degrees of freedom and $c_{I S_{1, E}, 1-\alpha}$ is the corresponding quantile of the null distribution of $I S_{1, E}$. The latter quantile will be determined by approximating the null distribution of $I S_{1, E}$ by a suitably scaled

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 |
| :---: | :---: | :---: | :---: | :---: |
| $I S_{1, E}$ | 0.012 | 0.010 | 0.011 | 0.011 |
| $I S_{1, M}$ | 0.011 | 0.009 | 0.011 | 0.011 |
| $I S_{2, E}$ | 0.023 | 0.012 | 0.015 | 0.010 |
| $I S_{2, M}$ | 0.023 | 0.011 | 0.013 | 0.009 |

and shifted $\chi^{2}$-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 $I S_{1, M}$ was independently introduced by Barabesi, Cerasa, Cerioli and Perrotta (2021).

Now, consider the second significant digit. Denote by $\vartheta=$ $\mathbb{E}\left(S(X) \mathbb{1}_{C_{2}(j)}(S(X))\right)=\frac{9}{10 \ln 10}$ the expectation of $S(X) \mathbb{1}_{C_{2}(j)}(S(X))$ if Benford is true. Let $\vartheta_{j}$ 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 $X_{i}$

$$
\operatorname{Sum}_{2, j}=\sum_{i=1}^{n} S\left(X_{i}\right) \mathbb{1}_{C_{2}(j)}\left(S\left(X_{i}\right)\right)
$$

Again, we have sums of $n$ independent identically distributed random variables $S\left(X_{i}\right) \mathbb{1}_{C_{2}(j)}\left(S\left(X_{i}\right)\right), i=1, \ldots, n$, and they have finite variance, we may assume that they are approximately normally distributed, and the standardized sums

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

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

Let be $\mathbf{R}_{2}=\left(R_{2,0}, \ldots, R_{2,9}\right)$ and let $\boldsymbol{\Sigma}_{R_{2}}$ 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

$$
I S_{2, E}=\mathbf{R}_{2}^{\prime} \mathbf{R}_{2} \quad \text { and } \quad I S_{2, M}=\mathbf{R}_{2}^{\prime} \boldsymbol{\Sigma}_{R_{2}}^{-1} \mathbf{R}_{2}
$$

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

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

|  |  | Data sets |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Test $\backslash n$ | Fibonacci <br> 1000 | Primes <br> 1000 | Population <br> 3998 | Share Prices <br> 369 |  |
| $I S_{1, E}$ | 1.00 | 0.00 | 0.00 | 0.89 |  |
| $I S_{1, M}$ | 1.00 | 0.00 | 0.00 | 0.88 |  |
| $I S_{2, E}$ | 1.00 | 0.02 | 0.00 | 0.26 |  |
| $I S_{2, M}$ | 1.00 | 0.00 | 0.00 | 0.33 |  |

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 $\alpha=0.01$.

## 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 that are due 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)[8].
$\# 3$ : Population $(n=3998)$
This data set consists of the number of inhabitants in cities worldwide that are larger than 100.000 people [23]. 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 [24]. Such data sets are assumed to behave like Benford, cf. 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 Figure 1. We see that the values of $R_{1, j}$ for the Share Prices and for the Fibonacci numbers are very close to zero indicating the Benford property. For the datasets Population and Prime Numbers the boxes are thick and far from zero indicating 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 $I S_{2, E}$ and $I S_{2, M}$. Table 6 contains the p-values for the tests $I S_{1, E}, I S_{1, M}, I S_{2, E}$, and $I S_{2, M}$.

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



Note that the values are rounded. This way, the entries especially for pvalues may become 1.00 or 0.00 . The p-value of (nearly) 1.00 of Fibonacci numbers indicates the well-known fact that they are nearly perfect Benford. The (rounded) p-value of 0.00 signals that prime numbers are not Benford. These two data sets, Fibonacci and Prime numbers, are selected for illustrating that all the tests considered yield a decision that confirms 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 confirm 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 confirmed by the three tests $I S_{1, E}, I S_{1, M}$ and $I S_{2, M}$, however, the test $I S_{2, E}$ does not reject Benford at the $\alpha=0.01$ level indicating that $I S_{2, E}$ has less power. The data set $\# 4$ (Share Prices) gives evidence of being Benford.

The results for all tests considered, KS1, KS2, GoF1, GoF2, MAD1, MAD2, $I S_{1, E}, I S_{1, M}, I S_{2, E}$ and $I S_{2, M}$, are presented in Table 7. Underlined values mean 'rejection', given $\alpha=0.01$. The classical tests essentially 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 less power. However, if we inccrease the sample size and take all prime numbers between 11 and 100,000 then Benford will be rejected also by all the tests based on the second significant digit.

## 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

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

| test | critical value | Data sets |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Population $\mathrm{n}=3998$ | Share Prices $\mathrm{n}=369$ | $\begin{gathered} \text { Fibonacci }^{1} \\ \mathrm{n}=1000 \end{gathered}$ | $\begin{aligned} & \text { Primes }^{1} \\ & \mathrm{n}=1000 \end{aligned}$ |
| KS1 | $1.42^{2}$ | 15.4 | 0.36 | 0.03 | $\underline{5.41}$ |
| KS2 | 1.46 | $\underline{5.00}$ | 0.58 | 0.20 | 1.41 |
| GoF1 | 20.09 | $\underline{1090}$ | 3.45 | 0.17 | $\underline{299.9}$ |
| 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 |
| $I S_{1, E}$ | 22.64 | $\underline{1274}$ | 4.10 | 0.18 | 332.8 |
| $I S_{1, M}$ | 21.67 | $\underline{1293}$ | 4.51 | 0.34 | $\underline{303.2}$ |
| $I S_{2, E}$ | 23.30 | $\underline{308}$ | 12.1 | 0.69 | 21.6 |
| $I S_{2, M}$ | 23.11 | $\underline{918}$ | 11.3 | 0.65 | 54.7 |

${ }^{1}$ for the tests KS2, GoF2, MAD2, $I S_{2, E}$, and $I S_{2, M}$ we removed all entries with only one digit.
2 the critical value for the KS1-test is obtained by Morrow (2014)
appealing as they apply the full significant of the data. Therefore the Invariant Sum tests use the full information in the data.

We have shown that all the tests give correct results for data sets for which there is theory whether the Benford property is true or not, cf. Tables 7 and 8. The last line in Table 9 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 \#4 are evidently Benford, the other two are not.

Table 8 p-values of the various test statistics

| test | Data sets |  |  |  |
| ---: | :---: | :---: | :---: | :---: |
|  | Population <br> $\mathrm{n}=3998$ | Share Prices <br> $\mathrm{n}=369$ | Fibonacci $^{1}$ <br> $\mathrm{n}=1000$ | Primes $^{1}$ <br> $\mathrm{n}=1000$ |
| 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 |
| $I S_{1, E}$ | 0.000 | 0.888 | 1.000 | 0.000 |
| $I S_{1, M}$ | 0.000 | 0.875 | 1.000 | 0.000 |
| $I S_{2, E}$ | 0.000 | 0.261 | 1.000 | 0.017 |
| $I S_{2, M}$ | 0.000 | 0.334 | 1.000 | 0.000 |
| BON | 0.000 | 1.000 | 1.000 | 0.000 |

[^0]Table A1 Variances of $S(X) \mathbb{1}_{C\left(d_{1}\right)}(S(X))$ 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 A2 Correlation matrix $\boldsymbol{\Sigma}_{1, R}$ of $\operatorname{Sum}_{1, j}$ of the Invariant sum tests $I S_{1, E}$ and $I S_{1, M}$ 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. |

## Statements and Declarations

There are no conflicts of interest.

## 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 $d_{1} \in\{1,2, \ldots, 9\}$ be given, then we have

$$
\begin{aligned}
\mathbb{E}\left(S(X) \mathbb{1}_{C\left(d_{1}\right)}(S(X))\right) & =\int_{d_{1}}^{d_{1}+1} t \cdot \frac{1}{t \ln 10} d t=\frac{1}{\ln 10} \\
\mathbb{E}\left(S(X) \mathbb{1}_{C\left(d_{1}\right)}(S(X))\right)^{2} & =\int_{d_{1}}^{d_{1}+1} t^{2} \cdot \frac{1}{t \ln 10} d t=\frac{2 d_{1}+1}{2 \ln 10} \\
\operatorname{var}\left(S(X) \mathbb{1}_{C\left(d_{1}\right)}(S(X))\right) & =\frac{2 d_{1}+1}{2 \ln 10}-\left(\frac{1}{\ln 10}\right)^{2} \\
\operatorname{cov}\left(S(X) \mathbb{1}_{C\left(d_{1}\right)}(S(X)), S(X) \mathbb{1}_{C\left(d_{1}^{\prime}\right)}(S(X))\right) & =-\frac{1}{(\ln 10)^{2}} \quad \text { if } \quad d_{1} \neq d_{1}^{\prime}
\end{aligned}
$$

which are already well-known results, cf. Barabesi, Cerasa, Cerioli and Perrotta (2021).

The variances of $S(X) \mathbb{1}_{C\left(d_{1}\right)}(S(X))$ are tabulated in table A1 and the correlation matrix $\boldsymbol{\Sigma}_{1, R}$ of the vector $\mathbf{R}_{1}$ is tabulated in table A2.

Table A3 Variances of $S(X) \mathbb{1}_{C_{2}\left(d_{2}\right)}(S(X))$ 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 |

Now, let $d_{2} \in\{0,1,2, \ldots, 9\}$ be given. Recall the sets $C_{2}\left(d_{2}\right)=\{x \in[1,10)$ : $\left.D_{2}(x)=d_{2}\right\}$. Then we have

$$
\begin{aligned}
\mathbb{E}\left(S(X) \mathbb{1}_{C_{2}\left(d_{2}\right)}(S(X))\right) & =\sum_{d_{1}=1}^{9} \int_{d_{1}+\frac{d_{2}}{10}}^{d_{1}+\frac{d_{2}+1}{10}} t \cdot \frac{1}{t \ln 10} d t \\
& =\frac{1}{\ln 10} \sum_{d_{1}=1}^{9}\left(d_{1}+\frac{d_{2}+1}{10}-\left(d_{1}+\frac{d_{2}}{10}\right)\right)=\frac{9}{10 \ln 10} \\
\mathbb{E}\left(S(X) \mathbb{1}_{C_{2}\left(d_{2}\right)}(S(X))\right)^{2} & =\sum_{d_{1}=1}^{9} \int_{d_{1}+\frac{d_{2}}{10}}^{d_{1}+\frac{d_{2}+1}{10}} t^{2} \cdot \frac{1}{t \ln 10} d t \\
& =\frac{1}{2 \ln 10} \sum_{d_{1}=1}^{9}\left(\left(d_{1}+\frac{d_{2}+1}{10}\right)^{2}-\left(d_{1}+\frac{d_{2}}{10}\right)^{2}\right) \\
& =\frac{1}{2 \ln 10} \sum_{d_{1}=1}^{9}\left(\frac{2}{10}\left(d_{1}+\frac{d_{2}}{10}\right)+\frac{1}{10^{2}}\right. \\
& =\frac{1}{200 \ln 10} \sum_{d_{1}=1}^{9}\left(20 \cdot d_{1}+2 d_{2}+1\right)=\frac{9\left(101+2 d_{2}\right)}{200 \ln 10} \\
\operatorname{var}\left(S(X) \mathbb{1}_{C_{2}\left(d_{2}\right)}(S(X))\right) & =\frac{9\left(101+2 d_{2}\right)}{200 \ln 10}-\left(\frac{9}{10 \ln 10}\right)^{2}
\end{aligned}
$$

The variances of $S(X) \mathbb{1}_{C_{2}\left(d_{2}\right)}(S(X))$ are tabulated in table A3.
To obtain the covariance note that the sets $C_{2}\left(d_{2}\right)$ and $C_{2}\left(d_{2}^{\prime}\right)$ are disjunct if $d_{2} \neq d_{2}^{\prime}$, and therefore

$$
\mathbb{E}\left(\left(S(X) \mathbb{1}_{C_{2}\left(d_{2}\right)}(S(X)) \cdot\left(S(X) \mathbb{1}_{C_{2}\left(d_{2}^{\prime}\right)}(S(X))\right)=0 \quad \text { if } \quad d_{2} \neq d_{2}^{\prime}\right.\right.
$$

Therefore the covariance equals

$$
\operatorname{cov}\left(S(X) \mathbb{1}_{C_{2}\left(d_{2}\right)}(S(X)), S(X) \mathbb{1}_{C_{2}\left(d_{2}^{\prime}\right)}(S(X))\right)=-\frac{81}{(10 \ln 10)^{2}} \quad \text { if } \quad d_{2} \neq d_{2}^{\prime}
$$

Therefore the correlation matrix $\boldsymbol{\Sigma}_{R, 2}$ of the vector $\mathbf{R}_{2}$ can be computed, see table A4.

Table A4 Correlation matrix $\boldsymbol{\Sigma}_{2, R}$ of $\operatorname{Sum}_{2, j}$ of the Invariant sum tests $I S_{2, E}$ and $I S_{2, M}$ 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 $I S_{1, E}$ and $I S_{1, M}$ let $\mathbf{U}_{1}$ be the matrix of Eigenvectors of the asymptotic correlation matrix $\boldsymbol{\Sigma}_{1, R}$. Let $\boldsymbol{\Lambda}_{1}=$ $\operatorname{diag}\left(\lambda_{1,1}, \ldots, \lambda_{1,9}\right)$, where the $\lambda_{1, j}, j=1, \ldots, 9$, are the Eigenvalues of $\boldsymbol{\Sigma}_{1, R}$.

Consider the random vector

$$
\mathbf{W}_{1}^{*}=\mathbf{\Lambda}_{1}^{-\frac{1}{2}} \mathbf{U}_{1}^{\prime} \mathbf{R}_{1}
$$

Obviously, $\operatorname{cov}\left(\mathbf{W}_{1}^{*}\right)=\boldsymbol{\Lambda}_{1}^{-\frac{1}{2}} \mathbf{U}_{1}^{\prime} \boldsymbol{\Sigma}_{1, R} \mathbf{U}_{1} \boldsymbol{\Lambda}_{1}^{-\frac{1}{2}}=\mathbf{I}_{1}$, where $\mathbf{I}_{1}$ is the $(9 \times 9)$ identity matrix. Let $\mathbf{0}_{1}$ be the null vector of dimension 9 . Therefore $\mathbf{W}_{1}^{*} \sim \mathcal{N}\left(\mathbf{0}_{1}, \mathbf{I}_{1}\right)$, asymptotically, under $H_{1,0}$. This way we have
$I S_{1, E}=\mathbf{R}_{1}^{\prime} \mathbf{R}_{1}=\mathbf{R}_{1}^{\prime} \mathbf{U}_{1} \boldsymbol{\Lambda}_{1}^{-\frac{1}{2}} \boldsymbol{\Lambda}_{1} \boldsymbol{\Lambda}_{1}^{-\frac{1}{2}} \mathbf{U}_{1}^{\prime} \mathbf{R}_{1}=\mathbf{W}_{1}^{*^{\prime}} \boldsymbol{\Lambda}_{1} \mathbf{W}_{1}^{*}=\sum_{j=1}^{9} \lambda_{1, j} W_{1, j}^{2}$
$I S_{1, M}=\mathbf{R}_{1}^{\prime} \boldsymbol{\Sigma}_{1, R}^{-1} \mathbf{R}_{1}=\mathbf{R}_{1}^{\prime} \mathbf{U}_{1} \boldsymbol{\Lambda}_{1}^{-1} \mathbf{U}_{1}^{\prime} \mathbf{R}_{1}=\mathbf{R}_{1}^{\prime} \mathbf{U}_{1} \boldsymbol{\Lambda}_{1, R}^{-1 / 2} \boldsymbol{\Lambda}_{1}^{-1 / 2} \mathbf{U}_{1}^{\prime} \mathbf{R}_{1}=\mathbf{W}_{1}^{*^{\prime}} \mathbf{W}_{1}^{*}=\sum_{j=1}^{9} W_{1, j}^{2}$
where the $W_{1, j}$ are the components of the vectors $\mathbf{W}_{1}^{*}$. Therefore the statistics $I S_{1, E}$ are, under $H_{0,1}$, asymptotically weighted sums of independent $\chi_{1}^{2}$ distributed random variables, where the weights $\lambda_{1, j}$ are the Eigenvalues of $\boldsymbol{\Sigma}_{1, R}$. Since the statistics $I S_{1, M}$ are asymptotically sums of nine squares of independent standard normal random variables, we have $I S_{1, M} \sim \chi_{9}^{2}$.

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

Consider the random vector

$$
\mathbf{W}_{2}^{*}=\mathbf{\Lambda}_{2}^{-\frac{1}{2}} \mathbf{U}_{2}^{\prime} \mathbf{R}_{2}
$$

Table C5 Eigenvalues of the correlation matrix $\boldsymbol{\Sigma}_{1, R}$ in the Invariant Sum Tests $I S_{1, E}$ ans $I S_{1, M}$

| 1.329 | 1.181 | 1.125 | 1.096 | 1.078 | 1.066 | 1.057 | 1.050 | 0.019 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |

Table C6 Eigenvalues of the correlation matrix $\boldsymbol{\Sigma}_{2, R}$ in the Invariant Sum Tests $I S_{2, E}$ ans $I S_{2, M}$

| 1.082 | 1.080 | 1.078 | 1.077 | 1.075 | 1.074 | 1.072 | 1.071 | 1.069 | 0.323 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |

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

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

where the $W_{2, j}$ are the components of the vectors $\mathbf{W}_{2}^{*}$. Therefore the statistics $I S_{2, E}$ are, under $H_{0,2}$ asymptotically weighted sums of independent $\chi_{1}^{2}$ distributed random variables, where the weights $\lambda_{2, j}$ are the Eigenvalues of $\boldsymbol{\Sigma}_{2, R}$. Since the statistics $I S_{2, M}$ are asymptotically sums of ten squares of independent standard normal random variables, we have $I S_{2, M} \sim \chi_{10}^{2}$.

## Appendix C

Approximation of the weighted sums by a $\chi^{2}$ distributed random variable. The quadratic forms $\mathbf{R}_{k}^{\prime} \mathbf{R}_{k}, \quad k=1,2$, will be approximated by (possibly noncentral) $\chi^{2}$ distributed random variables $Z_{k}$ suitably shifted and scaled according to the idea of Liu, Tang and Zhang (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 $\lambda_{1, j}, j=1, \ldots, 9$ are the Eigenvalues of the correlation matrix $\boldsymbol{\Sigma}_{1, R}$. The Eigenvalues can be found in Table C5. Analogously, recall that $\lambda_{2, j}, j=0, \ldots, 9$ are the Eigenvalues of the correlation matrix $\boldsymbol{\Sigma}_{2, R}$. The Eigenvalues can be found in Table C6.

Denote

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

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

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

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, Tang and Zhang (2009) we obtain that the noncentrality parameter of the $\chi^{2}$ approximation is zero, and the degrees of freedom $d f_{1}$, and the regression coefficients $\beta_{1,0}$ and $\beta_{1,1}$ are

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

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

$$
I S_{1, E}=\mathbf{R}_{1}^{\prime} \mathbf{R}_{1} \approx \beta_{1,1} Z_{1}+\beta_{1,0}, \quad \text { where } \quad Z_{1} \sim \chi_{d f_{1}}^{2}
$$

Therefore, if we choose $\alpha=0.01$, the critical value of the test $I S_{1, E}$ is $d_{c r i t, I S_{1, E}}=\beta_{1,1} \chi_{1-\alpha, d f_{1}}^{2}+\beta_{1,0}=22.6435$, which is close to the critical value $d_{\text {crit }, I S_{1, M}}=\chi_{0.99,9}^{2}=21.666$ of the $I S_{1, M}$-test.

In the case of the second significant digit $(\mathrm{k}=2)$ denote

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

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

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

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

$$
I S_{2, E}=\mathbf{R}_{2}^{\prime} \mathbf{R}_{2} \approx \beta_{2,1} Z_{2}+\beta_{2,0}, \quad \text { where } \quad Z_{2} \sim \chi_{d f_{2}}^{2}
$$

Therefore, if we choose $\alpha=0.01$, the critical value of the test $I S_{2, E}$ is approximately $d_{\text {crit }, I S_{2, E}}=\beta_{2,1} \chi_{1-\alpha, d f_{2}}^{2}+\beta_{2,0}=23.2963$, which is very close to the critical value $d_{\text {crit }, I S_{2, M}}=\chi_{0.99,10}^{2}=23.2093$ of the $I S_{2, M}$-test.

## Appendix D

On the (non-existing) Benford property for conditional distributions conditioned under $X>t$ with large $t$ and with small probability mass $\boldsymbol{P}(\boldsymbol{X}>\boldsymbol{t})$. 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-Balkemade Haan theorem, cf. Pickands (1975, theorem 7) or Balkema and de Haan (1974). Let $x=t+y, y \geq 0$.

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

$$
G P D(y ; k, \sigma)=\left\{\begin{array}{lll}
1-e^{-\frac{y}{\sigma}} & \text { if } & k=0 \\
1-\left(1-\frac{k y}{\sigma}\right)^{\frac{1}{k}} & \text { if } & k \neq 0
\end{array}\right.
$$

where $k$ is the shape parameter and $\sigma$ 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 $\sigma$ are given by the extreme value theory, cf. e.g. Falk (1987) or Kössler (1999). Note that the parameter $-k$ is sometimes called the extreme value index of the underlying distribution, cf. de Haan and Ferreira (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 $x \rightarrow \infty$. That is why we may assume that the parameter $k<0$. Then we have $\sigma=-k t$, cf. e.g. Falk (1987) or Kössler (1999). The conditional cdf $F(x \mid t)=F(t+y \mid t)=F_{t}(y)$ is approximated by

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

which is a Pareto cdf with scale parameter $t$ and shape parameter $\gamma:=\frac{1}{k}$. To obtain the probability $\left.P\left(D_{1}(X)=1\right) \mid X>t\right)$ 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

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

Looking at the shape of the function $g(\gamma), \gamma<0$ we see that for small values of $\gamma$ the probability $P\left(D_{1}(X)=1 \mid X>t\right)$ 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 $P\left(D_{1}(X)=1\right) \approx 0.428$ 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 $P\left(D_{1}(X)=1 \mid X>t\right) \gg 0.301$.

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

$$
\begin{aligned}
P\left(D_{2}(X)=l \mid X>t\right) & \approx \sum_{j=-\infty}^{\infty} \sum_{n=1}^{9}\left(G\left((n+1) \cdot\left(10^{j}+l\right)\right)-G\left(n \cdot\left(10^{j}+l\right)\right)\right) \\
& =\sum_{j=-1}^{\infty} \sum_{n=1}^{9}\left(\left(10^{j}(10 n+l)\right)^{\gamma}-\left(10^{j}(10 n+l+1)^{\gamma}\right)\right. \\
& =\sum_{n=1}^{9}\left((10 n+l)^{\gamma}-(10 n+l+1)^{\gamma}\right) \sum_{j=-1}^{\infty}\left(10^{\gamma}\right)^{j} \\
& =\sum_{n=1}^{9}\left((10 n+l)^{\gamma}-(10 n+l+1)^{\gamma}\right) \frac{1}{10^{\gamma} \cdot\left(1-10^{\gamma}\right)}
\end{aligned}
$$

In Table D 7 the probabilities $P\left(D_{2}(X)=l \mid X>t\right), l=0, \ldots, 9$ are presented for various values of the parameter $\gamma$ 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.

Table D7 Probabilities $P\left(D_{2}(X)=l \mid X>t\right), l=0, \ldots, 9$ for various values of $\gamma$ of the Pareto distribution

| $\gamma$ | $k$ | 1 |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | 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 |

## 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.

Table E8 Sample sizes $n$ for first digit, frequencies of the first significant digit, and values of the GoF1 statistic

|  | Data set |  | first significant digit |  |  |  |  |  |  |  |  | GoF1 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | n | 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 E9 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 |

## References

[1] Nigrini, M.J.: Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection. John Wiley \& Sons, Hoboken, New Jersey (2012)
[2] Kössler, W., H.-J. Lenz, Wang, X.D.: Is the Benford law useful for data quality assesment? In: Knoth, S., Schmid, W. (eds.) Frontiers in Statistical Quality Control, vol. 13, pp. 391-406. Springer, New York (2021)
[3] Cerqueti, R., Lupi, C.: Some new tests of conformity with Benford's law. Stats 4, 745-761 (2021)
[4] Newcomb, S.: Note on the frequency of use of the different digits in natural numbers. American Journal of Mathematics 4 (1), 39-40 (1881)
[5] Benford, F.: The law of anomalous numbers. Proceedings of the American Philosophical Sciety 78 (4), 551-572 (1938)
[6] Berger, A., Hill, T.P.: A basic theory of Benford's law. Probability Surveys 8, 1-126 (2011)
[7] Nigrini, M.J.: The detection of income evasion through an analysis of digital distributions. PhD thesis, University of Cincinnati (1992)
[8] Diekmann, A.: Not the first digit! using Benford's law to detect fraudulent scientific data. Journal of Applied Statistics 34, 321-329 (2007)
[9] Berger, A., Hill, T.P.: An Introduction to Benford's Law. Princeton University Press, Princeton and Oxford (2015)
[10] Pinkham, R.S.: On the distribution of first significant digits. Ann. Math. Statistics 32 4, 1223-1230 (1961)
[11] Pearson, K.: 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 (1900)
[12] Hein, J., Zobrist, R., Konrad, C., Schüpfer, G.: Scientific fraud in 20 falsified anesthesia papers. Aneasthesist 61(61), 543-549 (2012)
[13] Mebane, W.R.: Election Fraud or Strategic Voting? Can Seconddigit Tests Tell the Difference? https://citeseerx.ist.psu.edu/viewdoc/ download?doi=10.1.1.697.3403\&rep=rep1\&type=pdf (2010)
[14] Kolmogorov, A.N.: Sulla determinazione empirica di una legge di distribuzione. Giorn. dell'Inst Ital. degli An. 4, 83-91 (1933)
[15] Smirnov, N.V.: Tables for estimating goodness of fit of empirical distribution. Annals of Mathematical Statistics 19(2), 279-281 (1948)
[16] Darling, D.A.: The Kolmogorov, Cramer-von-Mises test. Annals of Mathematical Statistics 28(4), 823-838 (1957)
[17] Miller, L.H.: Table of percentage points of Kolmogorov statistics. Journal of the American Statstical Association 51(273), 111-121 (1956)
[18] Morrow, J.: Benford's law, families of distributions and a test basis. Discussion Paper No 1291, Centre for Economic Performance, LSE, London (2014)
[19] Kazemitabar, J., Kazemitabar, J.: Benford test based on logarithmic property. International Journal of Auditing Technology 4, 279-291 (2022)
[20] Cerqueti, R., Maggi, M.: Data validity and statistical conformity with Benford's law. Chaos, Solutions and Fractals 144 (2021)
[21] Allaart, P.C.: An invariant sum characterization of Benford's law. Journal of Applied Probability 34 (1), 288-291 (1997)
[22] Barabesi, L., Cerasa, A., Cerioli, A., Perrotta, D.: On characterizations and tests of Benford's law. Journal of the American Statistical Association (2021)
[23] UNStats Report https://unstats.un.org/unsd/demographic-social/ products/dyb/documents/dyb2016/table08.pdf. "[Online; accessed ]" (2016)
[24] Der Tagesspiegel: So war der Tag: DAX verließen die Kräfte, No.24476, 13.3.2021, p. 18
[25] Liu, H., Tang, Y., Zhang, H.H.: A new chi-square approximation to the distribution of nonnegative definite quadratic forms in noncentral normal variables. Computational Statistics \& Data Analysis 53, 853-856 (2009)
[26] Pickands, J.: Statistical inference using extreme order statistics. Annals of Statistics 3, 119-135 (1975)
[27] Balkema, A., de Haan, L.: Residual life time at great age. Annals of Probability 2, 792-804 (1974)
[28] Smith, R.L.: Estimating tails of probability distributions. Annals of Statistics 15, 1174-1207 (1987)
[29] Falk, M.: Best attainable rate of joint convergence of extremes. In:

Extreme Value Theory, Ed. J. Hüsler, R.-D. Reiss, Proceedings of a Conference Held in Oberwolfach, Dec.6-12, 1987, pp. 1-9. Springer, New York (1989)
[30] Kössler, W.: A new one-sided variable inspection plan for continuous distribution functions. Allgemeines Statistisches Archiv 83, 416-433 (1999)
[31] de Haan, L., Fereira, A.: Extreme Value Theory. Springer, New York (2006)


[^0]:    ${ }^{2}$ for the tests KS2, GoF2, MAD2, $I S_{2, E}$, and $I S_{2, M}$ we removed all entries with only one digit.

