Severe testing of Benford’s law

Roy Cerqueti¹,² · Claudio Lupi³

Received: 20 January 2023 / Accepted: 5 February 2023 / Published online: 28 February 2023
© The Author(s) under exclusive licence to Sociedad de Estadística e Investigación Operativa 2023

Abstract
Benford’s law is often used to support critical decisions related to data quality or the presence of data manipulations or even fraud in large datasets. However, many authors argue that conventional statistical tests will reject the null of data “Benford-ness” if applied in samples of the typical size in this kind of applications, even in the presence of tiny and practically unimportant deviations from Benford’s law. Therefore, they suggest using alternative criteria that, however, lack solid statistical foundations. This paper contributes to the debate on the “large n” (or “excess power”) problem in the context of Benford’s law testing. This issue is discussed in relation with the notion of severity testing for goodness-of-fit tests, with a specific focus on tests for conformity with Benford’s law. To do so, we also derive the asymptotic distribution of the mean absolute deviation (MAD) statistic as well as an asymptotic standard normal test. Finally, the severity testing principle is applied to six controversial large datasets to assess their “Benford-ness”.

Keywords Benford’s law · Data quality · Fraud discovery · Goodness-of-fit · Large n problem · Severity

Mathematics Subject Classification 62E20 · 62F03

Both authors contributed equally to this work.

Claudio Lupi
lupi@unimol.it

Roy Cerqueti
roy.cerqueti@uniroma1.it

¹ Department of Social and Economic Sciences, La Sapienza University of Rome, P.le Aldo Moro 5, 00185 Rome, Italy
² GRANEM, Université d’Angers, SFR Confluences, 49000 Angers, France
³ Department of Economics, University of Molise, Via F. De Sanctis, 86100 Campobasso, Italy
1 Introduction

Benford’s law, so named after Benford (1938), predicts that in many real instances the relative frequency of the first significant digits in a set of numbers approaches

\[ b_i = \Pr(X = d_i) = \log_{10} \left(1 + \frac{1}{d_i}\right) \]  

(1)

with \( d_i \in \{1, \ldots, 9\} \) for the first digit case or \( d_i \in \{10, \ldots, 99\} \) for the first-two digits case.

In fact, Benford’s law can be considered an instance of Stigler’s law of eponimy (Stigler 1980) since it was first discovered, nearly sixty years before Benford, by the American astronomer Simon Newcomb who noted:

“That the ten digits do not occur with equal frequency must be evident to any one making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones. The first significant figure is oftener 1 than any other digit, and the frequency diminishes up to 9. [...] The law of probability of the occurrence of numbers is such that all mantissae of their logarithms are equally probable.” (Newcomb 1881, emphasis in original)

Benford (1938) expanded Newcomb’s original idea, and offered a number of examples from different disciplines where the law seemed to hold. Later, Benford’s law was successfully applied in many different fields, spanning, e.g. accounting, astrophysics, finance, epidemiology, geography, and even musicology.

The reasons why data from many disciplines seem to obey Benford’s law has been deeply investigated and some explanations have been put forward (see, e.g. Raimi 1976; Hill 1995a, b; Rodriguez 2004; Fewster 2009; Block and Savits 2010; Ross 2011; Whyman et al. 2016). Based on this widespread regularity, Benford’s law is often used to investigate data quality and discover possible data manipulation or fraud in large datasets (see, e.g. Nigrini 2012; Kaiser 2019; Li et al. 2019). The idea is that, provided some minimal conditions are met, genuine data should obey Benford’s law and non-conformity with Benford’s law may indicate the presence of data problems. Such critical investigations should be based on robust and firmly grounded statistical evidence: however, this is not always the case.

From an empirical point of view, data conformity with Benford’s law can be tested using different criteria and statistical tests. Among the available tests, Pearson’s \( \chi^2 \) test has a predominant role. However, the use of the \( \chi^2 \) and other goodness-of-fit statistical tests has been forcefully criticized in a large part of the literature on Benford’s law on the grounds that they have “excessive power”—i.e. given a large enough sample size, they tend to reject the null of conformity with Benford’s law even in the presence of tiny and practically unimportant deviations (see, e.g. Nigrini 2012; Druică et al. 2018; Kossovsky 2021). For this reason, Drake and Nigrini (2000) and Nigrini (2012) favour the use of the mean absolute deviation (MAD) criterion, a measure of the average distance between the relative frequencies of the observed digits with Benford’s
Severe testing of Benford’s law

679

theoretical frequencies:

\[ \text{MAD} = \frac{1}{k} \sum_{i=1}^{k} |p_i - b_i| \]  

(2)

where \( k \) is the number of digits (\( k = 9 \) for the first digit case; \( k = 90 \) for the first-two digits case), \( p_i \) are the proportions of the first significant digits in the sample, and \( b_i \) are Benford’s probabilities from (1). According to its proposers, this quantity does not depend on the sample size \( n \), and therefore should not incur in the “excess power” problem (see, e.g. Nigrini 2012, p. 158). On the basis of personal experience and experimentation on many different datasets, Drake and Nigrini (2000) and Nigrini (2012, Table 7.1) propose some fixed thresholds of the MAD to classify data into four classes of conformity with Benford’s law corresponding to “close conformity”, “acceptable conformity”, “marginally acceptable conformity”, and “non-conformity”. These thresholds have been largely accepted in the applied literature and the MAD since then has become a workhorse of Benford’s analysis. In our view, basing sensitive decisions that may critically affect different aspects of society on criteria that lack solid statistical foundations is not a satisfactory answer to the large \( n \) (or “excess power”) problem raised in the literature.

The large \( n \) problem is very well known in the statistical literature (see, e.g. Berkson 1938; Lindley 1957; Cohen 1994; Granger 1998, to cite a few) and simply reflects the fact that any consistent test will asymptotically (with \( n \to \infty \)) reject with probability 1 any fixed false hypothesis, however close to the null this might be. In fact, this should not be considered a “problem” at all, but rather a desirable property of the test. Instead, the problem lies in not recognizing that given that the power of the test increases with the sample size \( n \), a rejection in the presence of a small \( n \) (low power) is stronger than one in the presence of a very large number of observations (high power). Giving the same weight to all rejections irrespective of \( n \) may bring to the fallacy of rejection in the sense of Mayo and Spanos (2006, 2011). What is of interest to us is the possibility of assessing the largest discrepancy from the null hypothesis warranted by the data at hand: then, depending on the specific context, we will be able to evaluate if that discrepancy is substantive or not.

In some circumstances, we might be interested in rejecting the null (in favour of a specific alternative) only in the presence of sizeable deviations from the null: this is certainly true in the case of tests of conformity with Benford’s law. The solution we propose is at the same time rigorous, elegant, and simple: we propose that the severity principle (see, e.g. Mayo and Spanos 2006, 2010, 2011; Mayo 2018) should be routinely applied when testing for goodness-of-fit in general, and for conformity with Benford’s distribution in particular. Severity allows us to focus on the presence (or absence) of what we may regard as a substantive discrepancy.

In this paper, we show how the post-data severity evaluation (PDSE) can be conveniently implemented in the special case of tests of conformity with Benford’s law. In our view, the severity principle finds a perfect application in goodness-of-fit testing in general, and in the assessment of Benford’s law in particular.

The structure of the paper is as follows: the next Section illustrates the severity principle. Section 3 describes how the severity principle can be applied to the case of
tests of Benford’s law. In Sect. 4, we apply the severity principle in testing Benford’s law on some controversial dataset to shed some light on their “Benford-ness”. Section 5 draws some conclusions.

2 The severity principle

Despite being potentially important in many real-world applications, the severity principle is still unfamiliar to most statisticians. The knowledge of this important concept has been relegated mostly among those interested in the relationships between statistics and the philosophy of science. Unsurprisingly, papers adopting the severe testing approach are virtually absent in the applied statistical literature.

Rejecting the null hypothesis on the basis of a small (or even very small) test $p$ value is not informative of the size of the discrepancy existing between the null and the data at hand. Very powerful tests (e.g. in the presence of a very large sample) may reject the null even for practically unimportant deviations.

Given the delicate decisions that can be taken on the basis of (non-)conformity with Benford’s law and the typically large sample sizes used to test Benford’s law, here we are mostly concerned with what Mayo and Spanos (2006) call the fallacy of rejection, where evidence against $H_0$ is misinterpreted as evidence for a particular $H_1$.

The idea of severity testing revolves around the intuition that rejection of the null hypothesis based on a test with low power for detecting a discrepancy $\gamma$ provides stronger evidence for the presence of a discrepancy $\gamma$ than rejection based on a substantially more powerful test. By its very nature, severity is based on a post-data perspective, when the direction of the possible discrepancy is known.

We are interested in the severity with which claim $C$ passes test $T$ with an outcome $T(x)$ (Mayo and Spanos 2006; Mayo 2018):

$$SEV[\text{test } T, \text{ outcome } T(x), \text{ claim } C]$$

which will be abbreviated simply as $SEV[T(x)]$ for ease of notation. For the claim $C$ to pass a severe test with data $x$, we require that

1. data $x$ agree with claim $C$, and
2. with high probability (at least $> 0.5$), test $T$ would have produced a result that accords less with claim $C$ if $C$ is false.

To explain the severity principle in relation with the sample size, we start from an example borrowed and adapted from Mayo and Spanos (2010).

Assume that we are interested in a one-sided test of the mean. The sample $X := (X_1, \ldots, X_n)$ is such that $X_i \sim NIID(\mu, \sigma^2)$ with known $\sigma^2 = 4$ (for simplicity). The test statistic is $T_n(X) := (\bar{X} - \mu_0)/\sigma\bar{X}$ where $\bar{X}$ is the sample mean and $\sigma\bar{X} = \sigma/\sqrt{n}$. The null and the alternative hypothesis of interest are $H_0 : \mu = \mu_0 = 0$ and $H_1 : \mu = \mu_1 = \mu_0 + \gamma > 0$, respectively. Under the null, $T_n(X) \sim N(0, 1)$.

Suppose that we test the null hypothesis on four observed samples $(x_1, x_2, x_3, \text{ and } x_4)$ of different size ($n_1 = 100, n_2 = 200, n_3 = 500, \text{ and } n_4 = 1000$). Suppose further that in each sample we obtain an identical result, $T_{n_i}(x_i) = 2 \forall i \in \{1, \ldots, 4\}$. Using
the usual significance levels, the four tests point to the rejection of the null hypothesis, with a $p$ value equal to 0.0228. However, this result is not informative about the size of the possible discrepancy with the null.

Assume that we are interested in the existence of a substantive discrepancy $\gamma$ that, in our specific context, we suppose that can be quantified as $\gamma > 0.2$. How robust is the inference that $\mu = \mu_0 + \gamma > 0.2$ in the four samples, given that the test statistic in each sample is equal to 2? We have to consider the severity of the claim $C : \mu > 0.2$, with a test outcome 2 using our test $T$. The fact that the test rejects the null is consistent with claim $C$. However, we need also that with high probability the test outcome should agree less with $C$ than the actual outcome does, if $C$ is false. This means that we must evaluate

$$SEV[T_n(X)] = \Pr[T_n(X) \leq T_n(x_o); \mu > \mu_1 \text{ is false}]$$

where $T_n(x_o)$ is the test statistic calculated on the observed sample and “;” means “calculated assuming” the specified condition. Of course, saying that $\mu > \mu_1$ is false is equivalent to saying that $\mu \leq \mu_1$ is true. In fact, it is sufficient to compute

$$SEV[T_n(X)] = \Pr[T_n(X) \leq T_n(x_o); \mu = \mu_1]$$

because for any $\mu < \mu_1$ we would obtain larger values of the severity.

The situation is summarized in Fig. 1. When $\mu = \mu_1 = \mu_0 + \gamma$ is assumed true, the distribution of $T_n(X)$ becomes $N[\sqrt{n} \gamma / \sigma, 1]$. The effect is that the distribution shifts to the right for higher values of $n$. Therefore, the rejection of the null implies different levels of severity (with respect to our alternative of interest) for different sample sizes: severity is higher for smaller values of $n$, ranging from 0.841 ($n = 100$) to 0.123 ($n = 1000$). That is, despite the fact that the null hypothesis is rejected at the same significance level, evidence in favour of our alternative of interest becomes weaker as $n$ grows.

Put under another perspective, the combined effect of discrepancy and sample size on severity can be visualized as in Fig. 2, where severity is plotted as a function of the discrepancy $\gamma$ for four different values of the sample size $n$. From Fig. 2, it is easy to see that the alternative hypothesis $H_1 : \gamma > 0.2$ is supported with decrease in severity as $n$ increases. The figure also highlights that different discrepancies can be supported with the same severity for different values of $n$. For example, for $n = 100$, the null $H_0 : \gamma = 0$ can be rejected with severity 0.9 in favour of the alternative $H_1 : \gamma > 0.144$, whereas in the presence of $n = 1000$ the null can be rejected with the same severity in favour of the alternative $H_1 : \gamma > 0.045$. Read in this way, Fig. 2 allows us to evaluate severity of rejection in a flexible way, leaving to the expert’s judgement the assessment of the practical importance of the rejection. In so doing, we are in line with the expert judgement approach for identifying the thresholds of close/acceptable/marginally acceptable conformity with Benford’s Law proposed, e.g. by Nigrini (2012) for the case of the MAD.
Fig. 1 Severity of the test in four samples of different size. The solid curve is the distribution of the test statistic under the null $\mu = \mu_0 = 0$; the dashed curve under the specific alternative $\mu = \mu_0 + \gamma = 0.2$. The vertical dashed line represents the value of the test statistic. The reported values of $SEV(2)$ represent the severity of the test in the different samples.

Fig. 2 Severity of the test as a function of discrepancy $\gamma$ in four samples of different size. The curves represent $Pr \left[ T_n(x) \leq T_n(x_0) \right]$ with $T_n(X) \sim N\left[ \sqrt{n} \gamma / \sigma, 1 \right]$ with $n = 100$ (solid), $n = 200$ (dashed), $n = 500$ (dotted), $n = 1000$ (dot-dashed). The dashed vertical line represents the hypothesized discrepancy of interest, $\gamma = 0.2$. The dashed horizontal line corresponds to a severity level equal to 0.9.
3 Severe tests of Benford’s law

3.1 Normal test

In order to maintain a direct relation with Nigrini’s $\text{MAD}$, it would be interesting to use a $\text{MAD}$-based test in evaluating data “Benford-ness”.

Cerqueti and Lupi (2021) show that, in the presence of a random sample, under the null hypothesis of data conformity with Benford’s law

$$\sqrt{n}\frac{|p_i - b_i|}{\sqrt{b_i(1-b_i)}} = \sqrt{n}\frac{|e_i|}{\sqrt{b_i(1-b_i)}} \overset{d}{\to} N\left(\frac{\sqrt{2}}{\pi}, 1 - \frac{2}{\pi}\right)$$ (6)

as $n \to \infty$, where of course $e_i = p_i - b_i$.

Equation (6) can be written in vector form as

$$\sqrt{n}D^{-1}|e| \overset{d}{\to} N\left(\frac{\sqrt{2}}{\pi}i, R\right)$$ (7)

where $D = \text{diag}\left(\sqrt{b_i(1-b_i)}\right)$ (with $i = 1, \ldots, k$), $e = (e_1, \ldots, e_k)'$, $i$ is a $k$-vector of 1s, and the covariance matrix $R$ is defined as (see Cerqueti and Lupi 2021)

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1k} \\ r_{12} & r_{22} & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1k} & r_{2k} & \cdots & r_{kk} \end{pmatrix} = \{r_{ij}\}$$ (8)

with

$$r_{ij} = \frac{2}{\pi} \left( \rho_{ij} \arcsin(\rho_{ij}) + \sqrt{1 - \rho_{ij}^2} \right) - \frac{2}{\pi}$$ (9)

where

$$\rho_{ij} = \begin{cases} -\sqrt{\frac{b_i b_j}{(1-b_i)(1-b_j)}} & \text{for } i \neq j \\ 1 & \text{for } i = j \end{cases}$$ (10)

It follows that

$$\sqrt{n}|e| \overset{d}{\to} N\left(\sqrt{\frac{2}{\pi}} D_1, DRD\right)$$ (11)

and

$$\sqrt{n}\text{MAD} = \frac{\sqrt{n}}{k}i'|e| \overset{d}{\to} N\left(\sqrt{\frac{2}{\pi k^2}} D_1, \frac{1}{k^2} D_1' DRD_1\right)$$ (12)

where, as usual, the prime denotes the transpose.

Equation (12) represents the asymptotic distribution of the $\text{MAD}$ under the null hypothesis that Benford’s law is valid. Finally, if Benford’s law holds, for $n$ large
Nigrini’s MAD is approximately distributed as \( N \left( \sqrt{\frac{2}{\pi n k^2}} t' D t, \frac{1}{nk^2} t' D R D t \right) \). Therefore, it is clear that Nigrini’s MAD, far from being independent of \( n \), under the null of conformity with Benford’s law is such that its expected value and standard deviation are inverse functions of \( \sqrt{n} \). To highlight that both the MAD and its expected value depend on \( n \) and \( k \), from now on we will denote them as \( \text{MAD}_{n,k} \) and \( E(\text{MAD}_{n,k}) \), respectively.

The use of the fixed thresholds proposed in Nigrini (2012, Table 7.1) implies that the \( \text{MAD}_{n,k} \) criterion will tend to reject too often for small values of \( n \) and to be too conservative for large ones.

Similarly to Barney and Schulzke (2016), our approach is to measure discrepancy in terms of the excess MAD

\[
\delta_{n,k} = \text{MAD}_{n,k} - E(\text{MAD}_{n,k}). \quad (13)
\]

Note that by (12), under the null hypothesis of data conformity with Benford’s law we have

\[
\tilde{\delta} = \frac{k\sqrt{n} \delta_{n,k}}{\sqrt{t' D R D t}} = \frac{k\sqrt{n} \left[ \text{MAD}_{n,k} - \sqrt{\frac{2}{\pi k^2}} t' D t \right]}{\sqrt{t' D R D t}} \xrightarrow{d} N(0, 1). \quad (14)
\]

In this way, we can obtain a standard normal test of conformity based on a scaled version of the excess MAD. We call this test “excess MAD test”. The null of conformity is \( H_0 : \delta_{n,k} = 0 \); the alternative of interest is one-sided, \( H_1 : \delta_{n,k} > 0 \) and can be specialized as \( H_1 : \delta_{n,k} > \delta^* > 0 \) to take into account substantive discrepancy. In fact, it should be stressed that the test \( p \) value is a post-data error probability for which we know the relevant tail: therefore we can legitimately disregard cases for which the test statistic is negative. In the present case, the distribution of the test statistic (14) under the alternative \( H_1 : \delta_{n,k} = \delta^* \) is approximately \( N \left( k\sqrt{n} \delta^*/\sqrt{t' D R D t}, 1 \right) \).

The distribution of the test statistic under the alternative can be used to carry out a post-data severity evaluation as in Sect. 2.

3.2 Chi-square test

Pearson’s \( \chi^2 \) goodness-of-fit test

\[
Q_n := n \sum_{i=1}^{k} \left( \frac{p_i - b_i}{b_i} \right)^2 \quad (15)
\]

is one of the most widely used tests to check data “Benford-ness”. Under the null that data conform to Benford’s law (i.e. \( p_i = b_i, \forall i = 1, \ldots, k \)), the test statistic is asymptotically distributed as \( \chi^2(k - 1) \), a central Chi-square random variable with
Severe testing of Benford’s law

$k - 1$ degrees of freedom, where $k = 9$ or $k = 90$ in the first digit or the first-two digits case, respectively. Under the alternative $p_i \neq b_i$. More precisely, assume that

$$p_i = b_i + \epsilon_i$$

with $\sum_{i=1}^{k} h_i = 0$: then for large $n$ the test statistic is approximately distributed as a non-central $\chi^2(k - 1, \psi)$ random variable, where the non-centrality parameter $\psi$ is (see, e.g. Lehmann and Romano 2005)

$$\psi = n \sum_{i=1}^{k} \frac{\epsilon_i^2}{b_i}. \quad (16)$$

The post-data severity evaluation exemplified in Sect. 2 for the normal case can be used also in the presence of a Chi-square statistical test, using the appropriate distributions. Again, it is up to the expert’s judgement the definition of what constitutes a practically relevant discrepancy from the null.

4 Application

In a recent paper, Kossovsky (2021, p. 426) argues that there are six datasets that are wrongly rejected as non-Benford by the Chi-square test, namely (i) Time in seconds between Global Earthquakes during 2012; (ii) USA Cities and Towns population; (iii) Canford PLC Price List; (iv) Star Distances from the Solar System; (v) Biological Genetic Measures; (vi) Oklahoma State positive expenses below $1 million in 2011.

In this Section, we apply the severity principle to the excess MAD test (14) and to the Chi-square test (15) to verify the possible existence of substantive discrepancies between the data and Benford’s law. Data are courtesy of Alex Kossovsky, and have been downloaded from https://web.williams.edu/Mathematics/sjmiller/public_html/benfordresources/.

4.1 Excess MAD test

Table 1 reports the results of the excess MAD normal test (14) for each of the six datasets under examination for both the first digit and the first-two digits case. Most of the tests point clearly towards the rejection of the null of conformity with Benford’s law, as highlighted by Kossovsky (2021) with reference to the Chi-square test, with the only exception of the first-two digits test applied to the USA city population dataset. However, the values of the $\text{MAD}_{n,k}$ criterion are always well below the non-conformity threshold, i.e. 0.015 for the first digit case and 0.0022 for the first-two digits case (Nigrini 2012, p. 160). In fact, Kossovsky (2021) argues that rejections with these datasets are just the manifestation of the “excess power” problem.
| Description                  | n   | First digit MADn, \( \delta \) | First-two digits MADn, \( \hat{\delta} \) | \( p \) value |
|------------------------------|-----|-----------------------------|---------------------------------|---------------|
| Earthquakes                 | 19,451 | 6.621                      | 0.00311                          | 0.00000       |
| USA city population         | 19,509 | 6.065                      | 0.00144                          | 0.00000       |
| Canford PLC price list      | 15,194 | 6.146                      | 0.00527                          | 0.00000       |
| Star distances              | 48,111 | 32.839                     | 0.00081                          | 0.00000       |
| Genetic measures            | 91,223 | 5.034                      | 0.00109                          | 0.00000       |
| Oklahoma State expenses     | 967,152 | 34.394                     | 0.00229                          | 0.00000       |
Figures 3 and 4 report the severity level of the tests on the six datasets for varying values of the excess $MAD$ for the first digit and the first-two digits case, respectively. In order to decide if the null hypothesis is severely rejected or not, we have to refer to some specific value of the excess $MAD$ that we may consider as evidence of a substantive deviation from Benford’s distribution. The exact value of this threshold should be left to the expert’s judgement. However, without necessarily endorsing Nigrini’s thresholds, here we note that the average distance between Nigrini’s fixed thresholds and the expected $MAD$ for $n$ varying over a large set of values is

$$\delta_k^* = \frac{1}{n_{\text{max}} - n_{\text{min}} + 1} \sum_{n=n_{\text{min}}}^{n_{\text{max}}} (t_k - E(MAD_{n,k})) \approx \begin{cases} 0.01 & \text{first digit} \\ 0.0012 & \text{first-two digits} \end{cases}$$

(17)

where $t_k$ ($k \in \{9, 90\}$) is Nigrini’s threshold, $E(MAD_{n,k})$ is the expected value of the $MAD$ for each specific number of observations ($n$) and digits ($k$), $n_{\text{max}} = 50,000$, and

Fig. 3 Severity of the excess MAD test on the first digit for Kossovsky’s data. a Earthquakes; b USA city population; c Canford PLC Price List; d Star distances; e Genetic measures; f Oklahoma State expenses
Fig. 4 Severity of the excess MAD test on the first-two digits for Kossovsky’s data. a Earthquakes; b USA city population; c Canford PLC Price List; d Star distances; e Genetic measures; f Oklahoma State expenses

$n_{\text{min}}$ is such to guarantee a minimum expected number of observations in each cell equal to 5 (i.e. $n_{\text{min}} = 110$ in the first digit case and $n_{\text{min}} = 1146$ in the first-two digits case). $t_k$ are taken as Nigrini’s thresholds for marginally acceptable conformity, i.e. 0.012 and 0.0018 for the first digit and the first-two digits, respectively. We will conventionally use the $\delta^*$ values resulting from (17) as the reference values to assess the existence of substantial departures from Benford’s law in our empirical exercise.

With reference to the first digit case, looking at Table 1 and Fig. 3, we note that, despite the fact that most of the $p$ values of the excess MAD test are tiny, nevertheless the largest discrepancy warranted by the data with high severity is always well below the “substantial discrepancy threshold” $\delta^*_9$. With reference to the first-two digits case, we find a discrepancy larger than $\delta^*_90$ with high severity only for the Oklahoma State expenses dataset (Fig. 4, panel F). A closer examination of this dataset reveals that there is a clear tendency towards “round” numbers: 10, 15, 20, 25, and so forth: see Fig. 5. In this case, Nigrini’s $MAD$ ($MAD_{n,90} = 0.00208$) suggests “marginally acceptable conformity”: we believe that this conclusion is totally misleading.
4.2 Chi-square test

We repeat the analysis of the six datasets using the standard Chi-square test (15). We again use Nigrini’s non-conformity thresholds as reference values. Simulating a large number of “perturbed” Benford’s distributions such that the MAD is equal to Nigrini’s threshold, it is possible to compute the corresponding average non-centrality parameter of the Chi-square distribution. The non-centrality parameters turn to be $0.0238 \times n$ and $0.0474 \times n$ for the first digit and the first-two digits case, respectively. The test statistics and the $p$ values are reported in Table 2, along with the values of the MAD criterion for ease of comparison. Again, the null is rejected for all the datasets, with the exception of the first-two digits distribution of the USA city population dataset.

Figures 6 and 7 plot severity of the Chi-square tests against the non-centrality parameter: to make plots more directly comparable, the non-centrality parameter is divided by $n$. As in the case of the excess MAD test, severity is high for fairly large discrepancies only in the case of the first-two digits test for the Oklahoma State expenses datasets (Fig. 7, panel F).

5 Concluding remarks

Benford’s law (after Benford 1938) is used in practice to support critical decisions in several different contexts, including assessing the possible existence of data manipulation or fraud. These decisions must rely on well-founded and trustworthy tests of data conformity with Benford’s law. However, many authors (see, e.g. Cho and Gaines 2007; Nigrini 2012; Tsagbey et al. 2017; Druică et al. 2018; Kossovsky 2021) have forcefully argued against using statistical tests to assess data conformity with Benford’s distribution, because of their “excess power” in the presence of large datasets. Alternative decision criteria, such as the mean absolute deviation (MAD; see, e.g.
Table 2  Chi-square tests of conformity with Benford’s law for Kossovsky’s data

| Description               | $n$   | First digit $\text{MAD}_{n,9}$ | Chi-square | $p$ value | First-two digits $\text{MAD}_{n,90}$ | Chi - square | $p$ value |
|---------------------------|-------|-------------------------------|------------|-----------|--------------------------------------|--------------|-----------|
| Earthquakes               | 19,451| 0.00479                       | 53.003     | 0.00000   | 0.00080                              | 146.066      | 0.00013   |
| USA city population       | 19,509| 0.00312                       | 17.524     | 0.02510   | 0.00062                              | 108.635      | 0.07712   |
| Canford PLC Price List    | 15,194| 0.00517                       | 57.739     | 0.00000   | 0.00149                              | 487.609      | 0.00000   |
| Star distances            | 48,111| 0.01088                       | 538.107    | 0.00000   | 0.00125                              | 862.868      | 0.00000   |
| Genetic measures          | 91,223| 0.00187                       | 34.686     | 0.00003   | 0.00035                              | 145.962      | 0.00013   |
| Oklahoma State expenses   | 967,152| 0.00253                      | 955.709    | 0.00000   | 0.00208                              | 79525.540    | 0.00000   |
Drake and Nigrini (2000; Nigrini 2012), have been proposed in the literature; however, we show that they lack firm statistical foundations.

This paper addresses the “excess power” (or “large \( n \)) controversy in the literature related to Benford’s law testing. We show that the post-data severity evaluation (Mayo and Spanos 2006, 2010, 2011; Mayo 2018) can and should be used to assess data “Benford-ness” in large data samples. In order to do so, we also derive the asymptotic distribution of Nigrini’s mean absolute deviation (MAD) statistic and propose an asymptotically standard normal test to which the severity principle can be easily applied, even using the information embodied in Nigrini’s judgmental thresholds. Finally, we carry out severe testing of Benford’s law on six controversial datasets (see Kossovsky 2021) using both the newly proposed normal test as well Pearson’s Chi-square test. With the exception of the Oklahoma State expenses dataset on which we disagree, we otherwise concur with Kossovsky (2021) that the discrepancies with Benford’s law in the majority of these datasets are not substantial. Rejections of the
null hypothesis clearly indicate the presence of some discrepancies, but their size may be regarded as practically unimportant.

Acknowledgements We would like to express our gratitude to Aris Spanos for his comments and suggestions on an early draft of this paper. Comments from Marcel Ausloos and two anonymous referees are gratefully acknowledged. We owe a special thank to Alex Kossovsky for having made public his data. All computations have been carried out using R 4.1.2 (R Development Core Team 2021); graphs greatly benefited from package “ggplot2” (Wickham 2016).

Funding No funds, grants, or other support was received.

Availability of data and materials The data used in the paper are publicly available at https://web.williams.edu/Mathematics/sjmiller/public_html/benfordresources/.

Code availability R scripts are available upon request.
Declarations

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Barney BJ, Schulzke KS (2016) Moderating “cry wolf” events with excess MAD in Benford’s law research and practice. J Forensic Account Res 1(1):A66–A90. https://doi.org/10.2308/jfar-51622
Benford F (1938) The law of anomalous numbers. Proc Am Philos Soc 78(4):551–572
Berkson J (1938) Some difficulties of interpretation encountered in the application of the chi-square test. J Am Stat Assoc 33(203):526–536. https://doi.org/10.1080/01621459.1938.10502329
Block HW, Savits TH (2010) A general example for Benford data. Am Stat 64(4):335–339. https://doi.org/10.1198/tast.2010.09169
Cerqueti R, Lupi C (2021) Some new tests of conformity with Benford’s law. Stats 4(3):745–761. https://doi.org/10.3390/stats4030044
Cho WKT, Gaines BJ (2007) Breaking the (Benford) law: statistical fraud detection in campaign finance. Am Stat 61(3):218–223. https://doi.org/10.1198/000313007x223496
Cohen J (1994) The earth is round (p<0.05). Am Psychol 49(12):997–1003. https://doi.org/10.1037/0003-066x.49.12.997
Drake PD, Nigrini MJ (2000) Computer assisted analytical procedures using Benford’s law. J Account Educ 18(2):127–146. https://doi.org/10.1016/s0748-5751(00)00008-7
Druică E, Oancea B, Vălsan C (2018) Benford’s law and the limits of digit analysis. Int J Account Inf Syst 31:75–82. https://doi.org/10.1016/j.accinf.2018.09.004
Fewster RM (2009) A simple explanation of Benford’s law. Am Stat 63(1):26–32. https://doi.org/10.1198/tast.2009.0005
Granger CW (1998) Extracting information from mega-panels and high-frequency data. Stat Neerl 52(3):258–272. https://doi.org/10.1111/1467-9574.00084
Hill TP (1995a) Base-invariance implies Benford’s law. Proc Am Math Soc 123(3):887–895. https://doi.org/10.1090/s0002-9939-1995-1233974-8
Hill TP (1995b) A statistical derivation of the significant-digit law. Stat Sci 10(4):354–363. https://doi.org/10.1214/ss/1177009869
Kaiser M (2019) Benford’s law as an indicator of survey reliability—can we trust our data? J Econ Surv 33(5):1602–1618. https://doi.org/10.1111/joes.12338
Kossovsky AE (2021) On the mistaken use of the chi-square test in Benford’s law. Stats 4(2):419–453. https://doi.org/10.3390/stats4020027
Lehmann EL, Romano JP (2005) Testing statistical hypotheses, 3rd edn. Springer Texts in Statistics. Springer, New York
Li F, Han S, Zhang H et al (2019) Application of Benford’s law in data analysis. J Phys Conf Ser 1168(3):032,133. https://doi.org/10.1088/1742-6596/1168/3/032133
Lindley DV (1957) A statistical paradox. Biometrika 44(1/2):187–192. https://doi.org/10.2307/2333251
Mayo DG (2018) Statistical inference as severe testing: how to get beyond the statistics wars. Cambridge University Press, Cambridge
Mayo DG, Spanos A (2006) Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. Br J Philos Sci 57(2):323–357. https://doi.org/10.1093/bjps/axl003
Mayo DG, Spanos A (2010) The error-statistical philosophy. In: Mayo DG, Spanos A (eds) Error and inference—recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science. Cambridge University Press, Cambridge, chap 2, pp 15–27
Mayo DG, Spanos A (2011) Error statistics. In: Bandyopadhyay PS, Forster MR (eds) Handbook of the philosophy of science, vol 7. Philosophy of statistics. Elsevier, pp 153–198. https://doi.org/10.1016/b978-0-444-51862-0.50005-8
Newcomb S (1881) Note on the frequency of use of the different digits in natural numbers. Am J Math 4(1):39–40. https://doi.org/10.2307/2369148
Nigrini MJ (2012) Benford’s law: applications for forensic accounting, auditing, and fraud detection. John Wiley & Sons, Hoboken. https://doi.org/10.1002/9781119203094
R Development Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://R-project.org
Raimi RA (1976) The first digit problem. Am Math Mon 83(7):521–538. https://doi.org/10.2307/2319349
Rodriguez RJ (2004) First significant digit patterns from mixtures of uniform distributions. Am Stat 58(1):64–71. https://doi.org/10.1198/0003130042782
Ross KA (2011) Benford’s law, a growth industry. Am Math Mon 118(7):571–583. https://doi.org/10.4169/amer.math.monthly.118.07.571
Stigler SM (1980) Stigler’s law of eponymy. Trans N Y Acad Sci 39(1 Series II):147–157. https://doi.org/10.1111/j.2164-0947.1980.tb02775.x
Tsagbey S, de Carvalho M, Page GL (2017) All data are wrong, but some are useful? Advocating the need for data auditing. Am Stat 71(3):231–235. https://doi.org/10.1080/00031305.2017.1311282
Whyman G, Shulzinger E, Bormashenko E (2016) Intuitive considerations clarifying the origin and applicability of the Benford law. Results Phys 6:3–6. https://doi.org/10.1016/j.rinp.2015.11.010
Wickham H (2016) ggplot2: Elegant graphics for data analysis. Use R!, Springer, New York. https://ggplot2.tidyverse.org

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.