SPANISH JOURNAL OF STATISTICS
Vol. **4** No. 1 2022, Pages 41–54
doi:https://doi.org/10.37830/SJS.2022.1.03

REGULAR ARTICLE

Testing Benford's law: from small to very large data sets

Leonardo Campanelli

All Saints University School of Medicine, leonardo.s.campanelli@gmail.com
Received: May 20, 2022, Accepted: December 30, 2022.

Abstract: We discuss some limitations of the use of generic tests, such as the Pearson's χ^2, for testing Benford's law. Statistics with known distribution and constructed under the specific null hypothesis that Benford's law holds, such as the Euclidean distance, are more appropriate when assessing the goodness-of-fit to Benford's law, and should be preferred over generic tests in quantitative analyses. The rule of thumb proposed by Goodman for compliance checking to Benford's law, instead, is shown to be statistically unfounded. For very large sample sizes ($N > 1000$), all existing statistical tests are inappropriate for testing Benford's law due to its empirical nature. We propose a new statistic whose sample values are asymptotically independent on the sample size making it a natural candidate for testing Benford's law in very large data sets.

Keywords: Benford's law, large data sets, goodness of fit, euclidean distance statistic, Pearson's χ^2

MSC: 62H15, 65C05, 62-07, 62G20

1 Introduction

Benford's law (Benford, 1938) on the distribution of the first significant digit (FSD) of numerical data is an empirical law that has been observed to emerge in disparate data sets, from finance (Nigrini, 1996; Cho and Gaines, 2007) and natural sciences (Sambridge et al., 2010) to COVID 19 data (Sambridge and Jackson, 2020; Campanelli, 2022).

By analyzing the data coming from very different distributions, such as length of rivers, populations of cities, etc., Benford found that the probability of occurrence of the first significant digit d, $P_B(d)$, followed the empirical law

$$\forall\, d \in \{1, ..., 9\}: \; P_B(d) = \log\left(1 + \frac{1}{d}\right). \tag{1}$$

Although we know today that Benford's law holds for some particular distributions (see Morrow (2014) and references therein) and that specific principles lead to the emergence of such a law (Hill, 1995a,b,c), the exact fundamental bases upon which Benford's law reposes are still unknown [for a review of Benford's law, see Miller (2015)].

The most common test in use for testing Benford's law is the Pearson's χ^2. In our case the χ^2 statistic can be written as

$$\chi^2 = N \sum_{d=1}^{9} \frac{[P_B(d) - P(d)]^2}{P_B(d)}, \tag{2}$$

where $P(d)$ is the observed relative frequency of the FSD d, and N is the sample size. However, such a test is based on the null hypothesis of a continuous distribution, and is generally conservative for testing discrete distributions as the Benford's one (Noether, 1963). This problem has been recently solved by Morrow (2014) who has computed asymptotically test values for this statistic under the specific null hypothesis that Benford's law holds.

Another estimator used for checking conformance to Benford's law is the "normalized Euclidean distance", d^*, introduced by Cho and Gaines (2007) and defined by

$$d^* = \frac{1}{D} \sqrt{\sum_{d=1}^{9} [P(d) - P_B(d)]^2}\,, \tag{3}$$

where $D = \sqrt{\sum_{d=1}^{8} P_B^2(d) + [P(9) - 1]^2}$ is a normalization factor that assures that d^* is bounded by 0 and 1. At the moment of its introduction, however, the properties of this new estimator were not well understood and no test values were reported. These problems have been solved by Morrow (2014), who has provided asymptotically test values for the "Euclidean distance"

$$d_N^* = \sqrt{N \sum_{d=1}^{9} [P(d) - P_B(d)]^2}\,, \tag{4}$$

and by the author who, recently enough (Campanelli, 2023), has found an empirical expression of its cumulative distribution function. A simple measure of fit to Benford's law, instead, has been proposed by Goodman (2016). His "rule of thumb" for conformance to Benford's law is $d^* \leq 0.25$.

One of the goals of this paper is to show the statistical incorrectness of Goodman's rule of thumb. Also, we will show that the use of p values of the χ^2 statistic for testing Benford's law is only appropriate for "qualitative" analyses, while the use of the Euclidean distance test should be preferred in "quantitative" analyses. Finally, we will discuss some limitations of existing statistical tests to assess the goodness-of-fit to Benford's law for very large number of data points ($N \leq 1000$) and/or small range of data, and we will propose a new statistic that overcome such limitations.

2 Euclidean distance statistic, χ^2 test, and Goodman's rule of thumb

The knowledge of the cumulative distribution function of the Euclidean distance statistic as a function of the sample size N, as derived in Campanelli (2023), makes possible the computation of p values and then allow us to check for the conformance of a set of data to Benford's law in a quantitative way. It is interesting, for example, to reconsider the data that allowed Benford to discover the law that now brings his name. In Table 1, we show the Euclidean distance d_N^* and its corresponding p value for the first-digit distribution of the twenty different groups of counts discussed by Benford in his original paper (Benford, 1938), while in Figure 1, we show the corresponding first-digit frequencies superimposed to Benford's law.

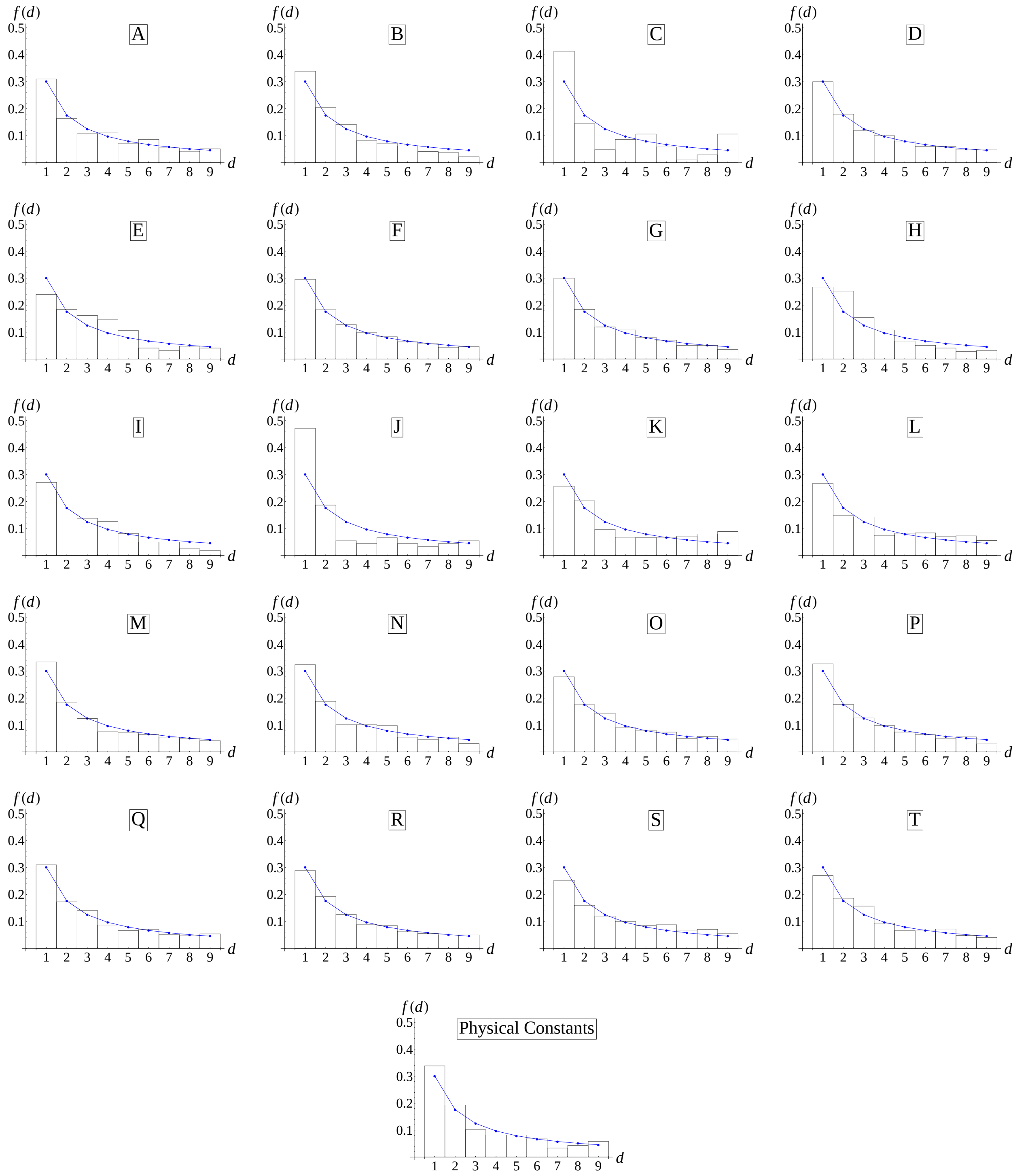

Figure 1: *Panels A to T.* Observed first-digit frequencies for the samples originally considered by Benford (1938) and shown in Table 1. *Bottom panel.* First-digit frequency of the values of the physical constants tabulated in Lide (2002). The (blue) continuous lines represent Benford's law.

Group	Title	N	d^*	d^*_N	p	χ^2	$p(\chi^2)$
A	Rivers, Area	335	0.0354	0.6705	0.78(9)	4.9617	0.7617
B	Population	3259	0.0602	3.5625	0.000(0)	118.63	0.0000
C	Constants	104	0.1581	1.6704	0.00(4)	24.441	0.0019
D	Newspapers	100	0.0107	0.1107	0.999(9)	0.1602	1.0000
E	Spec. Heat	1389	0.0949	3.6652	0.000(0)	111.21	0.0000
F	Pressure	703	0.0122	0.3360	0.99(6)	1.2704	0.9959
G	H.P. Lost	690	0.0188	0.5111	0.94(6)	3.4606	0.9022
H	Mol. Wgt.	1800	0.0931	4.0924	0.000(0)	125.76	0.0000
I	Drainage	159	0.0843	1.1018	0.17(8)	11.142	0.1938
J	Atomic Wgt.	91	0.1893	1.8718	0.000(9)	17.246	0.0277
K	n^{-1}, $\sqrt{n}$,...	5000	0.0827	6.0631	0.000(0)	440.76	0.0000
L	Design	560	0.0588	1.4430	0.02(4)	19.213	0.0138
M	$Digest$	308	0.0403	0.7333	0.69(8)	3.2271	0.9193
N	Cost Data	741	0.0443	1.2503	0.08(0)	15.601	0.0485
O	X-Ray Volts	707	0.0313	0.8622	0.48(9)	5.4256	0.7113
P	Am. League	1458	0.0315	1.2461	0.08(2)	14.595	0.0675
Q	Black Body	1165	0.0264	0.9350	0.37(7)	9.5229	0.3001
R	Addresses	342	0.0225	0.4314	0.98(0)	1.2966	0.9956
S	n^1,n^2,... $n!$	900	0.0583	1.8140	0.00(1)	24.994	0.0016
T	Death Rate	418	0.0480	1.0178	0.26(6)	7.5550	0.4781

Table 1: The Euclidean distance statistic d^*_N and its corresponding p value for the first-digit distribution of twenty different groups of counts discussed by Benford in his original paper (Benford, 1938). Also indicated are the total number of counts for each group, N, and the normalized Euclidean distance d^*. (Digits in parentheses at the third and fourth decimal places indicate an error on those digits of ±1). The last two columns show the χ^2 score and its corresponding p value, $p(\chi^2)$.

The data considered by Benford were collected from many different and disparate fields, from random numbers appearing within the covers of the same magazine to the values of physical constants [the rows K and S refer to an amalgamation of the observations of the first-digit frequencies of reciprocal and roots (row K) and powers and factorial (row S) of positive natural numbers]. At a first glance, the data suggest a certain regularity in the distribution of the first-digit, as it is evident in Figure 1, and as it was evident to Benford himself to the point that he claimed that "as no definite exceptions have ever been observed among true variables, the logarithmic law for large numbers evidently goes deeper among the roots of primal causes than our number system unaided can explain".

Surprisingly enough, however, half of the cases considered by Benford do not conform to Benford's law at a significance level of 0.10. Moreover, 40% do not conform at a significance level of 0.05 and one quarter do not conform at a significance level of 0.001.

The reasons for a non-conformance to Benford's law can be disparate. As stressed by Benford's himself (Benford, 1938), Benford's law "applies particularly to those outlaw numbers that are without known relationship rather than to those that individually follow an orderly course; and therefore the logarithmic relation is essentially a Law of Anomalous Numbers". Thus, groups E, H, J, and S do not comply to Benford's law probably because the underlying distributions of numbers do not satisfies Benford's requirement of "non-orderliness".

Another possibility is that the range of data is not sufficiently large to ensure conformance to Benford's law, which holds in the limit of an infinite range of data (Benford, 1938). This is probably the case of group C. In fact, if one considers the values of the physical constants as reported in Lide (2002), whose values extend on more than about 68 decades, one finds full conformance to Benford's law (see the bottom panel of Figure 1). In this case, we have $N = 207$, $d^* = 0.0550$, $d^*_{207} = 0.8196$, and $p = 0.55(8)$.

Groups B and K are the groups with the highest number of counts. A possible reason for the non-conformance in the case could be the enormous power of statistical tests for large N, which makes them too rigid to assess the goodness-of-fit well. This problem, and a possible solution, will be discussed in Sec. III.

It is worth observing that the use of the Cho-Gaines' normalized Euclidean distance d^* together with Goodman's rule of thumb for compliance to Benford' law, $d^* < 0.25$, would give a compliance to Benford's law for all groups of data in Table 1. Such a compliance is highly questionable. For example, consider group D and J. Both groups consist of a number of counts of about 100, but while D "seems" to follow Benford's law, J displays a big departure from it (see Figure 1).

Also, group B "seems" to display a high level of "Benfordness", but the use of the Euclidean distance statistic excludes the compliance to Benford's law at a significance level of 0.001.

This sort of "visual Benfordness" is then not reliable. This can be also understood by using the Pearson's χ^2 statistic. In the last two columns of Table 1, we show the value of the χ^2 and its corresponding p value, $p(\chi^2)$, for each group of count discussed by Benford (for the case of the values of the physical constants discussed above, we have $\chi^2 = 5.6983$, and $p(\chi^2) = 0.6810$). As it is clear, the p values of the χ^2 statistic differ, sometimes substantially, from the ones of the Euclidean distance statistic. This strongly indicates that the use of the χ^2 statistic for checking the conformance of a set of data to Benford's law is not completely reliable and should be used only for "qualitative" analyses. This is not surprising since, as it is well known, the χ^2 test has been designed for testing continuous distributions and is generally conservative for testing discrete ones, such as Benford's law (Noether, 1963).

In order to better understand the above two issues, the incorrectness of Goodman's rule of thumb and visual Benfordness, we have prepared six first-digit mock distributions with different total number of counts N. These are presented in Table 2 and visualized in Figure 2.

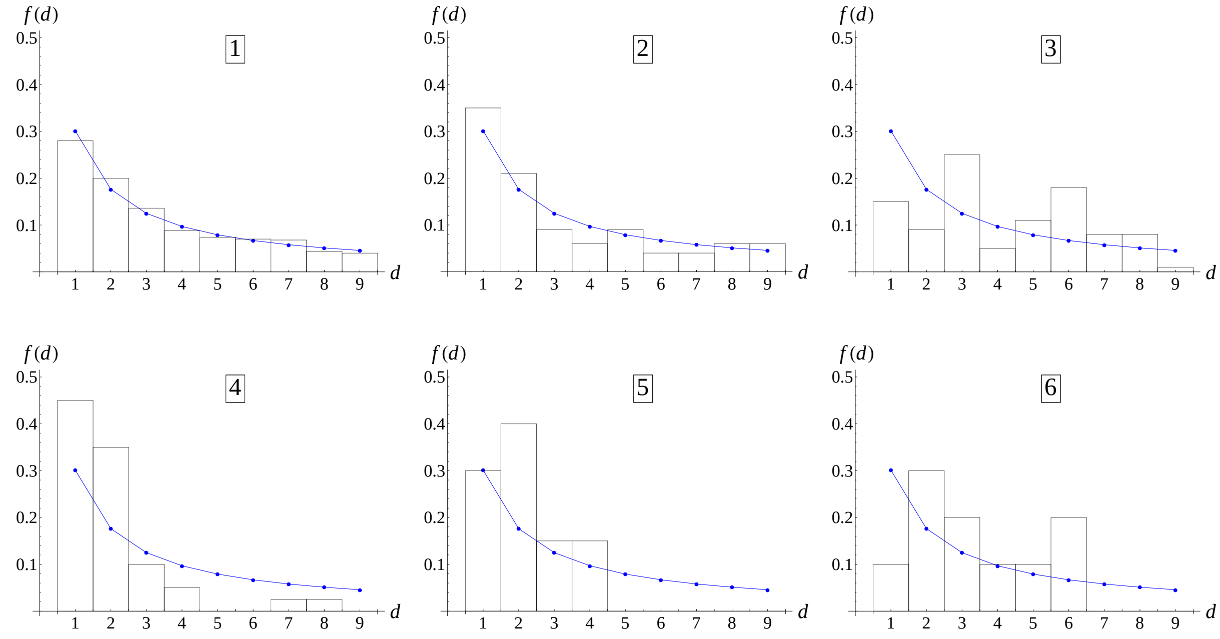

Figure 2: Graphical representations of the first-digit (mock) distributions in Table 2. The (blue) continuous lines represent Benford's law.

Group	N	$f(1)$	$f(2)$	$f(3)$	$f(4)$	$f(5)$	$f(6)$	$f(7)$	$f(8)$	$f(9)$	d^*	d_N^*	p	χ^2	$p(\chi^2)$
1	2500	0.280	0.200	0.136	0.088	0.074	0.070	0.068	0.044	0.040	0.0366	1.8949	0.000(7)	26.110	0.0010
2	500	0.360	0.200	0.100	0.060	0.080	0.040	0.040	0.060	0.060	0.0828	1.9190	0.000(6)	28.117	0.0005
3	100	0.150	0.090	0.250	0.050	0.110	0.180	0.080	0.008	0.010	0.2449	2.5375	0.000(0)	52.123	0.0000
4	40	0.450	0.350	0.100	0.050	0.000	0.000	0.025	0.025	0.000	0.2551	1.6718	0.00(4)	19.887	0.0108
5	20	0.300	0.400	0.150	0.150	0.000	0.000	0.000	0.000	0.000	0.2597	1.2034	0.10(2)	12.397	0.1343
6	10	0.100	0.300	0.200	0.100	0.100	0.200	0.000	0.000	0.000	0.2856	0.9361	$[0.25, 0.50]$	6.9145	0.5459

Table 2: The normalized Euclidean distance d^*, and the Euclidean distance statistic d_N^* and its corresponding p value, for six first-digit mock frequency distributions, $f(d)$, with different total number of counts N. (Digits in parentheses at the third and fourth decimal places indicate an error on those digits of ± 1). The last two columns show the χ^2 score and its corresponding p value, $p(\chi^2)$.

A look at Figure 2 would indicate high Benfordness of group 1, a moderate Benfordness of group 2, a low-level Benfordness of group 4, and non-Benfordness of groups 3, 5, and 6. The Euclidean distance statistic, on the contrary, shows that groups 1 and 2 do not conform to Benford's law at a significance level of 0.001 and group 4 at a significance level of 0.005. Moreover, groups 5 and 6 do conform to Benford's law at significance levels of 0.10 and > 0.25, respectively. Also, although groups 1, 2, and 3 do comply to Benford's law according to Goodman's rule, they do not at a significance level of 0.001 according to the Euclidean distance statistic. Finally, while the Benfordness of groups 5 and 6 should be rejected by Goodman's rule, the Euclidean distance statistic indicates a compliance to Benford's law at very high significance levels. Qualitatively, one can reach similar conclusions by using the χ^2 statistic (see the last two columns in Table 2).

Before considering group 4, it is worth noticing that Goodman's rule of thumb was obtained by the author by considering 40 empirical data sets displaying some visual "degree" of Benfordness (such a degree of Benfordness was not quantified by Goodman). He found that 95% of data sets had a d^* smaller than 0.256. So, he concluded that a value $d^* > 0.25$ is a strong indication of non-compliance to Benford's law, independently on the total number of counts N. We already showed that Goodman's rule generates wrong results when applied to data sets with counts larger or smaller than $N = 40$. But also when considering N around 40, one finds that Goodman's rule is not reliable. Indeed, let's now consider group 4 which contains exactly 40 counts. With a d^* value of 0.2551 and a $d^*_N = 1.6718$, the null hypothesis of conformance to Benford's law cannot be rejected at a significance level of 0.05 according to Goodman results, while it is rejected at a significance level of 0.005 by the Euclidean distance test. In this case, then, Goodman's rule is very conservative in rejecting the null hypothesis. This is probably due to the fact that the 40 empirical data sets used by Goodman had different "levels" of Benfordness.

3 The problem with very large data sets: ε-Benford's law

The Euclidean distance and the χ^2 tests, and in general all other tests used for checking the compliance of a data set to Benford's law, are very sensitive to the sample size N. In particular, they have enormous power for large N making them too rigid to assess the goodness-of-fit well: even a tiny deviation of the first-digit counts from Benford's distribution will be statistically significant. The severity of the existing tests for testing Benford's law for large N, can be traced back to the following reasons:

i) Benford's law does not represent a true law of numbers;

ii) Benford's law emerges in the limit of infinite range of the underlying distribution.

The emergence of Benford's law from a particular sample depends on the properties of the underlying distribution. However, no general criteria has be found that fully explain when and why Benford's law holds for a generic set of data. So, one major problem when testing for Benford's law is that it is not always possible to know in advance if a set of data is expected to follow it or not. This means that the rejection/acceptance of the null can be misleading when the underlying distribution is "close" to but not exactly Benford's, and this regardless of data quality. This problem is exacerbated by the increase of power of statistical tests with the sample size and, for very large sample sizes, say $N \gg 1000$, it makes any statistical test unreliable.

Also, Benford's law, even when it is known to hold exactly Morrow (2014); Hill (1995a,b,c), emerges from underlying distributions that extend on infinite ranges. In real applications, however, the set of data is restricted to a finite range and, typically, to just few decades. For finite ranges,

Year	N	d^*	d^*_N	χ^2	p
1994	9632	0.052	5.3	350	> 0.10
1996	11108	0.081	8.9	510	< 0.001
1998	9694	0.061	6.2	420	$(0.01, 0.05)$
2000	10771	0.072	7.7	670	$(0.001, 0.01)$
2002	10348	0.097	10	1100	< 0.001
2004	8396	0.130	12	2200	< 0.001

Table 3: The normalized Euclidean distance d^*, the Euclidean distance statistic d^*_N, and the χ^2 score for the first-digit distribution of in-kind contributions for six particular election cycles discussed by Cho and Gaines (2007). Also indicated is the total number of counts for each group, N. The last column shows the p values of the Euclidean distance statistic for a ε-Benford's distribution with $\varepsilon = 0.20$.

then, we expect a deviation from Benford's law even if the underlying distribution is exactly Benford. Moreover, we expect that such a deviation becomes statistically significant at large N.

In order to overcome the problem of the enormous power of existing statistical tests for large N, Cho and Gaines (2007) introduced the normalized Euclidean distance statistic in the attempt to quantify the deviation of a data set from Benford's law. However, as pointed out by the authors, the use of this statistic can only identify possible anomalies that deserve further inspection, but does not represent a "quantitative" statistical tool for testing Benford's law.

To better understand this point, let us consider the data analyzed by Cho and Gaines (2007) about the first-digit frequencies of in-kind contributions for six particular election cycles. In Table 3, we show the normalized Euclidean distance d^*, the Euclidean distance statistic d^*_N, and the χ^2 score for such distributions. Due to the extremely large values of both d^*_N and χ^2, the null hypothesis of conformance to Benford's law is rejected at any conceivable significance level for all years. The very large number of counts for each group, of order of 10^5, makes the Euclidean distance and χ^2 tests too powerful to properly assess the goodness-of-fit. However, the values of the normalized Euclidean statistic d^*, as well as those of the Euclidean statistic d^*_N, indicate that the last two elections exhibit a somewhat worse fit than their earlier counterparts (Cho and Gaines, 2007).

In the rest of this Section, we will extend the work of Cho and Gaines by making the identification of anomalies more "quantitative".

We first give the following definition. A random variable X, whose first-digit probability distribution function is $P_\varepsilon(d)$, follows a ε-Benford's distribution iff

$$\forall\, d \in \{1, ..., 9\}: \left| \frac{P_\varepsilon(d) - P_B(d)}{P_B(d)} \right| \leq \varepsilon. \tag{5}$$

Here, the positive parameter $\varepsilon \times 100\%$ quantifies the maximum percentage deviation of the values of the first-digit distribution of X from Benford's law. [Notice that if a random variable X follows a ε-Benford distribution, it automatically ε-satisfies Benford's law in the sense specified byMorrow (2014).]

The first-digit frequencies of in-kind contributions discussed above fail to conform to Benford's law even if their deviations from the law are relatively small, as confirmed by the smallness of the normalized Euclidean distance statistic. Indeed, the underlying random variable could be "intrinsically" ε-Benford, or it became so due to the limitedness of the range of data. Whatever is the case, we may assess the goodness-of-fit of such frequencies to ε-Benford's law after finding the appropriate test values of the Euclidean distance statistic for a ε-Benford distribution.

To this end we performed a Monte Carlo simulation consisting, for each sample size N, of n draws from a Benford's distribution $P_B(d)$, with each value of $P_B(d)$ being multiplied by a (pseudo-)random number in the interval $[1-\varepsilon, 1+\varepsilon]$, thus obtaining the ε-Benford distribution $P_\varepsilon(d)$. In particular, we considered the cases $\varepsilon = 0.05, 0.10, 0.15, 0.20, 0.25$, and we took $n = 10^5$ for $50 \leq N \leq 10000$ and $n = 10^4$ for $10000 < N \leq 100000$. We started with $N = 50$ and $N = 100$, and then we proceeded up to 1000 by steps of 100, up to 10000 by steps of 1000, and up to 100000 by steps of 10000. We then evaluated the Euclidean distance statistic for the ε-Benford distribution, $d_N^{(\varepsilon)}$, as

$$d_N^{(\varepsilon)} = \sqrt{N \sum_{d=1}^{9} [P_\varepsilon(d) - P_B(d)]^2}\,. \tag{6}$$

The observed probability distribution function of the Euclidean distance statistic exhibits a regular dependence of the sample size N. This is apparent in the upper panels of Figure 3, where we show its mean $\overline{d_N^{(\varepsilon)}}$ and its standard deviation $s_N^{(\varepsilon)}$ as a function of the sample size N. In the middle and lower panels of Figure 3, instead, we show the test values $d_{N,1-\alpha}^{(\varepsilon)}$ for $\alpha = 0.10, 0.05, 0.01$, and 0.001 [the test values $d_{N,1-\alpha}^{(\varepsilon)}$ are defined as Cdf $[d_{N,1-\alpha}^{(\varepsilon)}] = 1-\alpha$, where Cdf $[d_N^{(\varepsilon)}]$ is the (observed) cumulative distribution function of $d_N^{(\varepsilon)}$]. The (blue) continuous lines represent fits of the observed quantities and are divided in two intervals, $50 \leq N \leq 10^3$ and $10^3 \leq N \leq 10^5$. All nonlinear fits can be expressed as

$$\theta_N^{(\varepsilon)} = \left(a + bN^{-1/2} + cN^{-1}\right)\sqrt{N}, \tag{7}$$

where $\theta_N^{(\varepsilon)}$ represents any of the variables $\overline{d_N^{(\varepsilon)}}$, $s_N^{(\varepsilon)}$, and $d_{N,1-\alpha}^{(\varepsilon)}$. The fitting values a, b, and c, for both $50 \leq N \leq 10^3$ and $10^3 \leq N \leq 10^5$, are shown in Table 4.

The choice of the fitting function in Eq. (7) is suggested by the behaviour of the quantities $\theta_N^{(\varepsilon)}/\sqrt{N}$, which numerically are found to be slowly decreasing function of N approaching constant limiting values (see Figure 3). Indeed, assuming that the parameters a, b, and c remain constant for $N > 10^5$, it follows from Eq. (7) that all quantities $\theta_N^{(\varepsilon)}/\sqrt{N}$ approach asymptotic constant values for a given ε,

$$\lim_{N\to\infty} \frac{\theta_N^{(\varepsilon)}}{\sqrt{N}} = \theta^{(\varepsilon)}. \tag{8}$$

A linear fit of these values as a function of ε gives

$$\overline{d^{(\varepsilon)}} = -0.0011 + 0.1960\,\varepsilon, \tag{9}$$

$$s^{(\varepsilon)} = -0.0004 + 0.0596\,\varepsilon, \tag{10}$$

and

$$d_{0.90}^{(\varepsilon)} = -0.0017 + 0.2791\,\varepsilon, \tag{11}$$

$$d_{0.95}^{(\varepsilon)} = -0.0020 + 0.3033\,\varepsilon, \tag{12}$$

$$d_{0.99}^{(\varepsilon)} = -0.0024 + 0.3410\,\varepsilon, \tag{13}$$

$$d_{0.999}^{(\varepsilon)} = -0.0029 + 0.3717\,\varepsilon. \tag{14}$$

We show the limiting values $\theta^{(\varepsilon)}$ and their corresponding linear fits in Figure 4. It is worth noticing that these fits cannot be extrapolated down to $\varepsilon = 0$. Indeed, from the discussion in Campanelli (2023), we expect $\theta^{(\varepsilon)} \to 0$ in the limit $\varepsilon \to 0$.

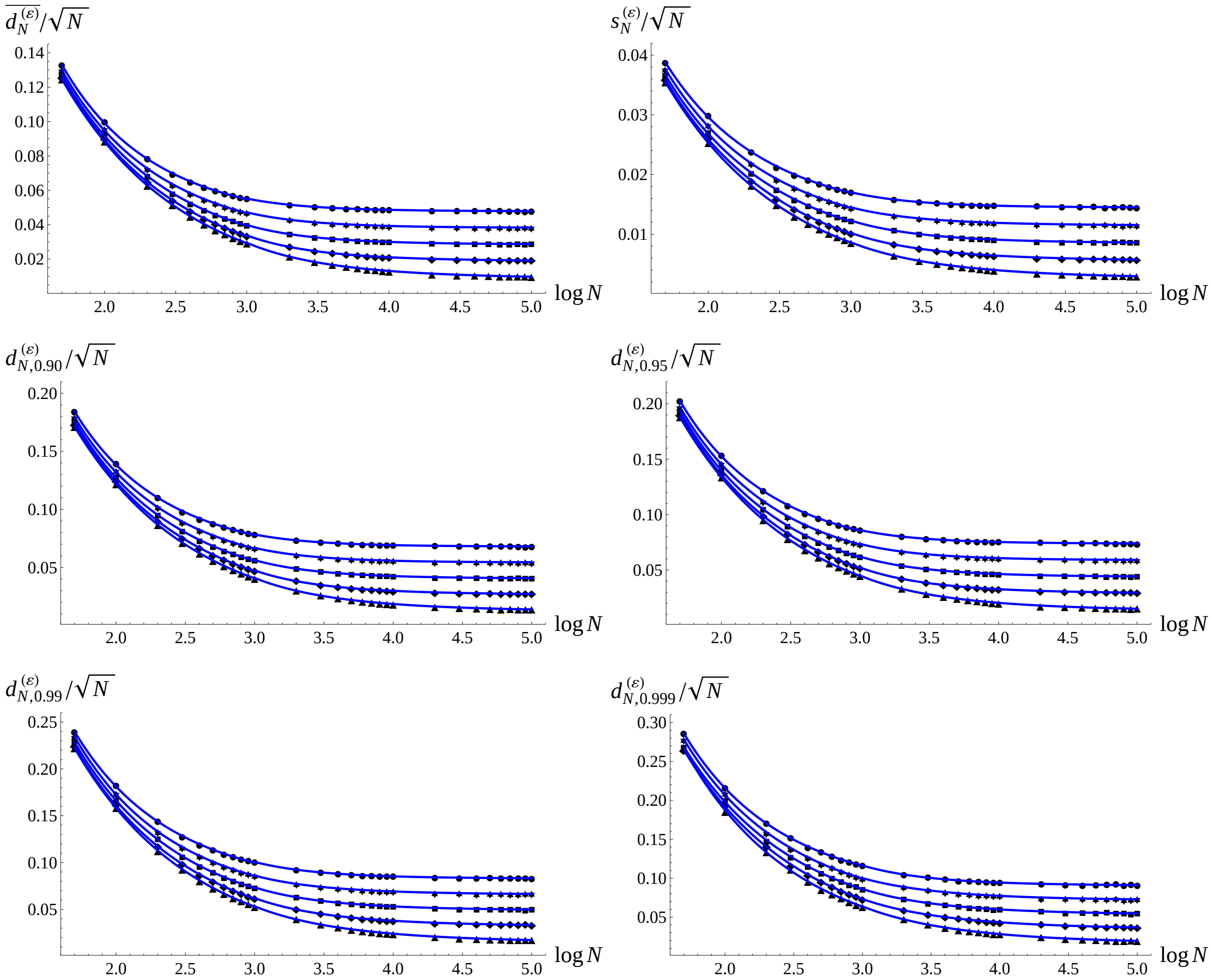

Figure 3: The mean (upper left panel), standard deviation (upper right panel), and test values (middle and lower panels) of the Euclidean distance statistic (6) as a function of the sample size N, together with their nonlinear fits (blue continuous lines), Eq. (7), for different values of ε. From bottom to top: $\varepsilon = 0.05, 0.10, 0.15, 0.20$, and 0.25.

$\overline{d_N^{(\varepsilon)}}$						
	$50 \leq N \leq 10^3$			$10^3 \leq N \leq 10^5$		
ε	a	b	c	a	b	c
0.05	0.0024	0.8429	0.1592	0.0085	0.3686	9.329
0.10	0.0090	0.7536	0.5061	0.0187	0.1252	11.10
0.15	0.0180	0.6478	0.8886	0.0285	0.0568	9.570
0.20	0.0287	0.5274	1.3311	0.0382	0.0327	7.999
0.25	0.0393	0.4382	1.5933	0.0478	0.0189	6.811
$s_N^{(\varepsilon)}$						
	$50 \leq N \leq 10^3$			$10^3 \leq N \leq 10^5$		
ε	a	b	c	a	b	c
0.05	0.0009	0.2425	0.0171	0.0025	0.1275	2.058
0.10	0.0031	0.2176	0.1027	0.0055	0.0648	2.637
0.15	0.0060	0.1875	0.2099	0.0085	0.0375	2.621
0.20	0.0091	0.1596	0.3028	0.0114	0.0259	2.231
0.25	0.0121	0.1423	0.3310	0.0145	0.0095	2.259
$d_{N,0.90}^{(\varepsilon)}$						
	$50 \leq N \leq 10^3$			$10^3 \leq N \leq 10^5$		
ε	a	b	c	a	b	c
0.05	0.0034	1.1701	0.1253	0.0120	0.5272	12.18
0.10	0.0132	1.0400	0.6570	0.0264	0.1943	14.82
0.15	0.0264	0.8842	1.2222	0.0403	0.0922	13.05
0.20	0.0412	0.7373	1.7073	0.0542	0.0480	11.12
0.25	0.0564	0.6109	2.0842	0.0679	0.0331	9.398
$d_{N,0.95}^{(\varepsilon)}$						
	$50 \leq N \leq 10^3$			$10^3 \leq N \leq 10^5$		
ε	a	b	c	a	b	c
0.05	0.0040	1.2794	0.1662	0.0130	0.6121	12.63
0.10	0.0151	1.1325	0.7568	0.0286	0.2607	15.38
0.15	0.0291	0.9790	1.2903	0.0437	0.1328	14.28
0.20	0.0452	0.8181	1.8199	0.0588	0.0742	12.41
0.25	0.0610	0.7041	2.1113	0.0737	0.0497	10.73
$d_{N,0.99}^{(\varepsilon)}$						
	$50 \leq N \leq 10^3$			$10^3 \leq N \leq 10^5$		
ε	a	b	c	a	b	c
0.05	0.0048	1.5143	0.1805	0.0145	0.8291	12.52
0.10	0.0180	1.3483	0.7872	0.0319	0.4373	16.38
0.15	0.0335	1.1962	1.3040	0.0488	0.2810	15.84
0.20	0.0520	1.0014	2.0509	0.0657	0.2074	13.80
0.25	0.0688	0.9084	2.1364	0.0828	0.0978	14.68
$d_{N,0.999}^{(\varepsilon)}$						
	$50 \leq N \leq 10^3$			$10^3 \leq N \leq 10^5$		
ε	a	b	c	a	b	c
0.05	0.0092	1.685	0.9586	0.0158	1.1156	12.73
0.10	0.0189	1.697	0.3014	0.0342	0.7717	14.63
0.15	0.0365	1.534	0.7803	0.0528	0.5464	16.07
0.20	0.0547	1.370	1.4935	0.0712	0.4710	13.99
0.25	0.0770	1.161	2.2538	0.0902	0.2273	18.90

Table 4: The values of the best-fitting parameters a, b, and c in Eq. (7) as a function of ε, for both $50 \leq N \leq 10^3$ and $10^3 \leq N \leq 10^5$.

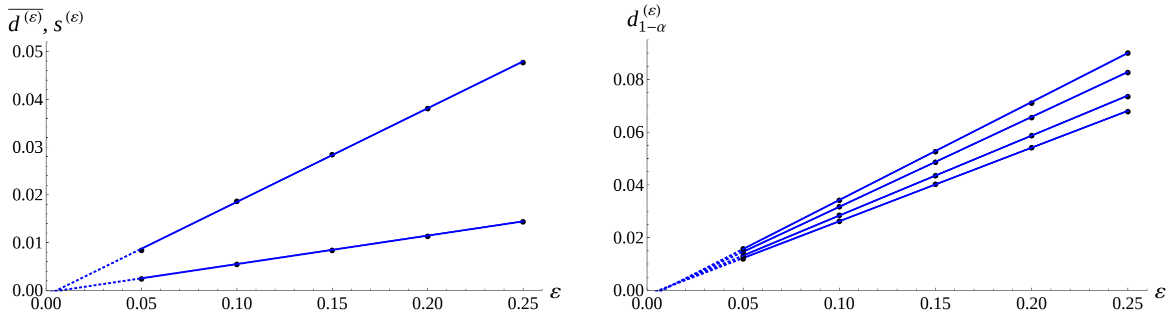

Figure 4: *Left panel.* The limiting values (8) of the mean (upper circles) and standard deviation (lower circles) of the Euclidean distance statistic (6) as a function of ε, together with their corresponding regression lines (blue continuous lines), Eqs. (9) and (10), respectively. *Right panel.* Limiting test-values (8) for the Euclidean distance statistic (6) as a function of ε, together with their corresponding regression lines (blue continuous lines), Eqs. (11)-(14). From bottom to top: $\alpha = 0.1, 0.05, 0.01$ and 0.001.

The mean, standard deviation, and test values of $d_N^{(\varepsilon)}$ grow as $\sqrt{N}$ for large N. Accordingly, the statistic $d_N^{(\varepsilon)}/\sqrt{N}$, whose sample values are by definition independent on N, is asymptotically independent on the sample size making its use a reliable tool for testing Benford's law in samples with large size. This statistic, then, solves the problem of the enormous power for large N of existing statistical tests.

Let us now re-consider the Cho and Gaines data discussed above. These data are divided in "homogeneous" groups, in the sense that the underlying statistical process for each group is the same, namely an in-kind contribution cataloged by the U.S.A. Federal Election Commission. As already noticed, these data sets do not comply to Benford's law (see Table 3). However, we can test the hypothesis that the data comply to a ε-Benford's distribution. We proceed as follows. We fix the value of ε by finding a group conforming to ε-Benford to a large significance level, let's say, bigger than 0.1. This is the case of the first election cycle (1994) for which $d_N^* = 5.3$ and $d_N^{(0.2)} = 5.5$. In other words, the data relative to the 1994 election conform to a 0.20-Benford's distribution at a significance level of 0.1. Accordingly, we can test the null (conformity to a 0.20-Benford's distribution) for the other cycles. The results are shown in the last column of Table 3. While the 1998 in-kind contribution conforms to a significance level of 0.05 and the 2000 one cannot be rejected at a significance level of 0.001, the years 1996, 2002, and 2004 present anomalies: the conformance of the data to a 0.20-Benford's distribution can be rejected at a significance level grater than 0.001. It is interesting to observe that, based on the normalized Euclidean distance statistic, Cho and Gaines (2007) found anomalies only for the years 2002 and 2004. The use of ε-Benford's distribution can be then used not only to identify but also to quantify possible anomalies in homogeneous sets of data.

The use of ε-Benford distributions can be also extended to the case of "non-homogeneous" data sets when the number of counts is large and/or the range of the data is small. An interesting application of Benford's law to physical and mathematical data sets was discussed by Sambridge et al. (2010). Their data are shown in Table 5. More than 50% of the data sets considered in their study has large number counts. In particular, groups S1, S3, S4, S6, S7, S8, S9, and S11 have values of N larger that 10^3 (groups S8 and S11 also have a small range of data, of order of 10^2). Not surprisingly, with the exception of S7 which well conforms to Benford's law, these large-number-count sets have huge values of d_N^* and χ^2 making the goodness-of-fit unreliable. However, they all comply to ε-Benford's

Set	Title	N	Range	d^*	d_N^*	p	χ^2	$p(\chi^2)$	ε
S1	Geomagnetic Field	36512	10^{10}	0.015	2.936	0.	49.90	0.	0.10
S2	Geomagnetic reversals	93	10^3	0.056	0.562	0.91(1)	3.608	0.8907	—
S3	Seis. wavespeeds below SW-Pacific	423776	10^6	0.009	6.041	0.	363.7	0.	0.04
S4	Earth's gravity	25917	10^9	0.035	5.829	0.	188.7	0.	0.15
S5	Exoplanet mass	401	10^5	0.056	1.163	0.13(0)	10.57	0.2274	—
S6	Pulsars rotation freq.	1861	10^4	0.060	2.699	0.	55.03	0.	0.25
S7	Fermi space tel. γ-ray source fluxes	1451	10^5	0.020	0.797	0.59(5)	12.56	0.1280	—
S8	Earthquake depths	248915	10^2	0.027	14.14	0.	1723	0.	0.11
S9	S-A seismogram	24000	10^5	0.030	4.840	0.	191.1	0.	0.15
S10	Green house gas em. by country	184	10^4	0.030	0.421	0.98(3)	2.049	0.9795	—
S11	Glob. Temp. anom. 1880-2008	1527	10^2	0.045	1.828	0.00(1)	34.61	0.	0.15
S12	Fund. Phys. constants	326	10^4	0.058	1.088	0.19(1)	9.615	0.2931	—
S13	Global Infectious disease cases	987	10^6	0.046	1.419	0.02(7)	15.00	0.0592	—
S14	Geometric series	1000	10^{21}	0.007	0.229	0.999(7)	0.417	0.9999	—
S15	Fibbonacci sequence	1000	10^{14}	0.003	0.091	1.	0.122	1.	—

Table 5: The normalized Euclidean distance d^*, the Euclidean distance statistic d_N^* with its corresponding p value, the χ^2 score with its corresponding p value, $p(\chi^2)$, of the first-digit distribution for various physical and mathematical data sets discussed by Sambridge et al. (2010). Also indicated are the total number of counts for each group, N, and the dynamic range of the data (max/min). The last column shows the value of ε such that the first-digit distribution of the counts conform to a ε-Benford's law at a significance level of $\alpha = 0.1$.

law at a significance level of $\alpha = 0.1$, with values of ε ranging from 0.04 to 0.25 (see the last column in Table 5).

4 Conclusions

Benford's law on the distribution of the first digits of numerical data sets has been observed to arise in multifarious classes of data, from natural sciences to finance. Compliance to Benford's law can be tested by using standard test statistics, such as the Pearson χ^2 statistic, the "Goodman's rule of thumb", and/or the recently introduced Euclidean distance statistic.

The main results of our analysis are as follows.

(i) For small and/or large number of data points, $N \leq 1000$, the use of p values of the χ^2 statistic for testing Benford's law is not completely reliable. This is because the χ^2 test, although being a very powerful tool in assessing the goodness-of-fit to any continuous distribution, it is generally conservative for testing discrete ones, like Benford's law. The χ^2 statistic should be then used only for "qualitative" analyses. For quantitative analyses, the Euclidean distance should be used, since the test based on this statistic has been explicitly constructed for testing Benford's law. The Goodman's rule of thumb, instead, should be always avoided when checking for the compliance to Benford's law. Its statistical groundlessness generates results in disagreement with both the χ^2 and the Euclidean distance tests.

(ii) We have discussed some limitations of statistical tests in assessing the goodness-of-fit to Benford's law for very large sample sizes ($N > 1000$) and/or very small ranges of data, and then proposed a possible solution to overcome such limitations. The solution comes from the observation that Benford's law is not in general a limiting distribution nor a fundamental law of numbers and then real distributions are often ε-Benford, in the sense that they deviate from Benford's law at a relative level of ε. Even a tiny deviation, however, may result in huge values of standard test statistics for large N, making any attempt to quantify the goodness-of-fit unfeasible. We have then considered a new statistic, the Euclidean distance statistic $d_N^{(\varepsilon)}$ for a ε-Benford distribution, and computed appropriate test values. The statistic $d_N^{(\varepsilon)}/\sqrt{N}$, whose sample values are by definition independent on N, is asymptotically independent on the sample size, making it a natural candidate for testing Benford's law in samples with very large size.

References

Benford, Frank (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 551–572.

Campanelli, Leonardo (2022). On the euclidean distance statistic of Benford's law. *Communications in Statistics-Theory and Methods*, 1–24.

Campanelli, Leonardo (2023). Breaking Benford's law: A statistical analysis of Covid-19 data using the euclidean distance statistic. *Statistics in Transition New Series*. In press.

Cho, Wendy K T and Brian J Gaines (2007). Breaking the (Benford) law: Statistical fraud detection in campaign finance. *The American Statistician 61*(3), 218–223.

Goodman, William (2016). The promises and pitfalls of Benford's law. *Significance 13*(3), 38–41.

Hill, Theodore P (1995a). Base-invariance implies Benford's law. *Proceedings of the American Mathematical Society 123*(3), 887–895.

Hill, Theodore P (1995b). The significant-digit phenomenon. *The American Mathematical Monthly 102*(4), 322–327.

Hill, Theodore P (1995c). A statistical derivation of the significant-digit law. *Statistical Science*, 354–363.

Lide, David R (2002). Handbook of chemistry and physics, eighty. CRC Press LLC, Internet Version of the 84th Edition.

Miller, S. J. (Ed.) (2015). *Benford's Law: Theory and Applications*. Princeton University Press, Princeton.

Morrow, John (2014). Benford's law, families of distributions and a test basis.

Nigrini, Mark J (1996). A taxpayer compliance application of Benford's law. *The Journal of the American Taxation Association 18*(1), 72.

Noether, Gottfried E (1963). Note on the Kolmogorov statistic in the discrete case. *Metrika 7*(1), 115–116.

Sambridge, Malcolm and Andrew Jackson (2020). National COVID numbers–Benford's law looks for errors. *Nature 581*(7809), 384–385.

Sambridge, Malcolm, Hrvoje Tkalčić, and A Jackson (2010). Benford's law in the natural sciences. *Geophysical Research Letters 37*(22).