Journal of Statistical Research *https://doi.org/10.3329/jsr.v58i2.80606*
2024, Vol. 58, No. 2, pp. 259-277 *ISSN 0256 - 422 X*

A TEST OF SIGNIFICANCE FOR BENFORD'S LAW BASED ON THE CHEBYSHEV DISTANCE

LEONARDO CAMPANELLI
All Saints University School of Medicine, 5145 Steeles Ave., M9L 1R5, Toronto, Canada, and
Mississauga Career College, 6341 Mississauga Rd., L5N 1A5, Mississauga, Canada
Email: leonardo.s.campanelli@gmail.com

SUMMARY

We show, by means of a numerical simulation, that the asymptotic ($n \geq 100$) cumulative distribution function of the Chebyshev distance statistic is well approximated by a log-normal function with parameters $\mu = -0.6183$ and $\sigma = 0.3561$ in the null hypothesis that Benford's law holds. The deviations of the cumulative function observed in Monte Carlo simulations from the empirical one are below 0.5%. This makes the statistical test based on the Chebyshev statistic accurate at a level of 1% when testing Benford's law for moderately large and large numbers of data points. Test values of the Chebyshev distance as a function of the sample size are also estimated empirically by performing a Monte Carlo simulation in the case of low n ($10 \leq n \leq 99$). The efficacy and power of the goodness-of-fit test based on the Chebyshev estimator are analyzed and compared with those based on the Pearson χ^2 and Kolmogorov-Smirnov statistics. Finally, an application of the Chebyshev test to the annual deaths counts by country is discussed.

Keywords and phrases: Goodness-of-Fit Test, Benford's Law, Mathematical Modelling, Chebyshev Distance, Max Statistic, World Death Counts, World Homicides Counts, World Deaths by Infectious Diseases, World Suicides Counts

AMS Classification: 65E20, 62G10, 62G30, 62P99

1 Introduction

Benford's law (Benford, 1938) is an empirical statistical law about the distribution of the first significant digit (FSD) d of numerical data sets that has been observed to emerge in many and disparate real-data sources.

Although the causes of appearance of this law in data are still not completely clear, the number of its applications has grown in recent years [for theoretical insights and general applications of Benford's law, see Miller (2015)]. Probably, the most famous applications are to detecting fraud in campaign finance (Cho and Gaines, 2007) and political elections (Roukema, 2013). Other interesting applications are in cryptology, where the law has been employed to examine the truthfulness of undeciphered numerical codes (Wase, 2021, Campanelli, 2022a), and in epidemiology, where Benford's law has been applied to the study of the temporal spread of infectious diseases, such as Covid

19 (Sambridge and Jackson, 2020, Farhadi, 2021, Campanelli, 2023) and Monkeypox (Campanelli, 2024a).

The most common tests in use for testing Benford's law in data are the Pearson's χ^2 and the Kolmogorov-Smirnov tests. These tests, however, have some limitations. The former has a low power for even moderately large sample sizes (Morrow, 2014), while the latter, being based on the hypothesis of a continuous distribution, is generally conservative for testing discrete distributions as the Benford's one (Noether, 1963). The latter problem has been recently solved by the author (Campanelli, 2024b) by showing that an appropriate linear transformation of the argument of the standard Kolmogorov cumulative distribution function makes the Kolmogorov-Smirnov test accurate at a level of 1% when testing Benford's law for moderately large and large numbers of data points ($n \geq 100$).

Other tests have been proposed in the literature to overcome the limitations of the χ^2 and Kolmogorov-Smirnov tests. They are based on new statistics, such as the "Chebyshev distance" statistic (also known as "max distance" statistic), introduced by Leemis et al. (2000), and the "normalized Euclidean distance" statistic, introduced by Cho and Gaines (2007). Asymptotic test values for these statistics have been subsequently provided by Morrow (2014).

For the case of the normalized Euclidean distance statistic, we have already expanded the work of Morrow (2014) by finding an empirical expression of its cumulative distribution function as a function of the sample size (Campanelli, 2022b, 2022c). The goal of this paper is twofold: to find an empirical expression of the asymptotic cumulative distribution function of the Chebyshev distance statistic, and to compare the efficacy and power of the test based on the Chebyshev estimator with those of the χ^2 and Kolmogorov-Smirnov tests.

2 Method

The Chebyshev distance statistic is defined by (Leemis et al., 2000, Morrow, 2014)

$$m_n = \sqrt{n} \max_d |f_B(d) - f_n(d)|, \tag{2.1}$$

where

$$f_B(d) = \log\left(1 + \frac{1}{d}\right) \tag{2.2}$$

is the probability mass function of a random Benford variable X, $f_B(d) = \text{Pr}(X \text{ has FSD} = d)$, and $f_n(d)$ is the observed first-digit distribution of a numerical set containing n data points.

In the left panel of Figure 1, we show the observed cumulative distribution function (Cdf) of the Chebyshev distance statistic, $F_{\text{obs}}(m_n)$, for $n = 500$ (black points), found by a Monte Carlo simulation based on $N = 10^5$ draws. [1]

[1] Monte Carlo simulations and fitting procedures were all performed by using *Mathematica* (Wolfram, 2023) with its build-in libraries, such as Mathematica's core randomness generator "RandomChoice" and "NonlinearModelFit". Notice also that special functions, such as the (elliptic) theta function $\vartheta_4(z, q)$ introduced in Section 7, are build-in functions in Mathematica and can be evaluated to an arbitrary numerical precision.

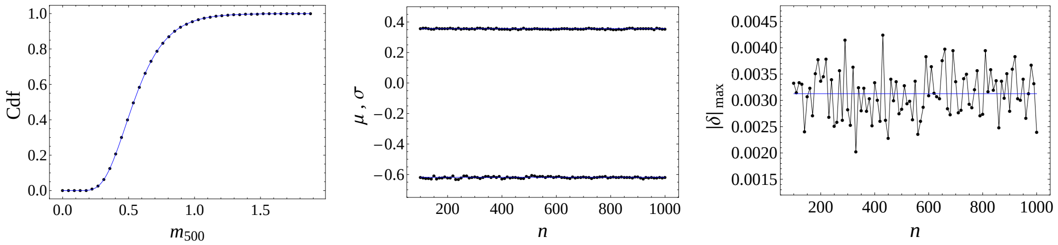

Figure 1: *Left panel.* The observed Cdf for $n = 500$ (black points), together with the best-fitted Cdf $F_{\text{log-normal}}(m_n)$ in Equation 2.3 (continuous blue line). *Middle panel.* The best-fit values of the parameters μ and σ in Equation 2.3 as a function of n. The (blue) continuous lines are the corresponding linear fits whose values are reported in Equation 3.1. *Right panel.* The maximum $|\delta|_{\max} = \max_{m_n} |\delta(m_n)|$ of the absolute value of the difference $\delta(m_n) = F_{\text{log-normal}}(m_n) - F_{\text{obs}}(m_n)$ between the theoretical Cdf with parameters in Equation 3.1 and the observed one as a function of n, together with its linear fit (the blue continuous line).

The (blue) continuous line in the panel represents the distribution obtained by fitting the observed one with the Cdf of a (standard) log-normal distribution with location parameter μ and shape parameter σ,

$$F_{\text{log-normal}}(m_n) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\ln m_n - \mu}{\sigma\sqrt{2}}\right)\right], \tag{2.3}$$

where $\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z dt\, e^{-t^2}$ is the error function (Abramowitz and Stegun, 1972). The values of the best-fit parameters are $\mu = -0.6184$ and $\sigma = 0.3554$. Since the absolute value of the difference between the best-fit cumulative function and the observed one has a maximum value as low as 0.0028, we repeated this procedure for each sample size n. The minimum and maximum n considered are 100 and 1000, respectively, and we proceeded with increments of $\Delta n = 10$.

3 Results

The best-fit values of the parameters μ and σ as a function of n are shown in the middle panel of Figure 1. The (blue) continuous lines are the fit lines

$$\mu = -0.6183, \quad \sigma = 0.3561. \tag{3.1}$$

In the right panel of Figure 1, we show $|\delta|_{\max} = \max_{m_n} |\delta(m_n)|$ as a function of n, namely the maximum of the absolute value of the differences $\delta(m_n) = F_{\text{log-normal}}(m_n) - F_{\text{obs}}(m_n)$ between the empirical Cdf with parameters in Equation 3.1 and the observed one, for each run with a given n. The average value of $|\delta|_{\max}$, represented by the (blue) continuous line, is about 0.0031.

The choice of the log-normal distribution is satisfactory, in the sense that it produces values of $|\delta|_{\max}$ comparable to the typical value of the statistical fluctuations associated to the finite number of draws N, which is expected to be about $1/\sqrt{N} \sim few \times 10^{-3}$. Moreover, since $|\delta|_{\max} < 0.005$

Table 1: Asymptotic ($n \geq 100$) test values $m_{n,1-\alpha}$, mean $\overline{m_n}$, standard deviation s_n, and skewness g_n of the Chebyshev distance statistic m_n in Equation 2.1 obtained by using the empirical Cdf in Equation 2.3 with parameters μ and σ given by Equation 3.1, and the empirical Cdf in Equation 4.1 with parameters α and β given by Equation 4.2. Also indicated are the test values obtained by Morrow (2014).

	$m_{n,0.90}$	$m_{n,0.95}$	$m_{n,0.99}$	$\overline{m_n}$	s_n	g_n
log-normal	0.85	0.97	1.23	0.57	0.21	1.15
beta prime	0.85	0.96	1.21	0.58	0.21	1.04
Morrow	0.851	0.967	1.212	–	–	–

for all n, our empirical Cdf can be used to calculate p values, and accordingly to test Benford's law, with an accuracy of 1%.

In Table 1, we show the test values $m_{n,0.90}$, $m_{n,0.95}$, and $m_{n,0.99}$, [2] the sample mean $\overline{m_n}$, the standard deviation s_n, and the skewness g_n of the Chebyshev distance statistic in Equation 2.1 obtained by using the Cdf in Equation 2.3 with parameters μ and σ given by Equation 3.1. The asymptotic test values for $\alpha = 0.10, 0.05$, and 0.01 are in agreement with those obtained by Morrow (2014) by a different statistical procedure based on a Monte Carlo simulation with $N = 10^6$ draws and sample sizes up to $n = 500$.

4 Other Choices for the Cdf of the Chebyshev Distance

As we showed before, the log-normal distribution provides a reasonably good fit to the Cdf of the Chebyshev distance. Other distributions, however, could in principle give better fits. If we restrict ourself to the case of "known" unimodal distribution i) with positive support, ii) positively skewed and, for simplicity, iii) having at most two parameters (μ and σ for the case of the log-normal distribution), then the log-normal distribution gives, to the best of our knowledge, the closest fit to the observed Cdf of the Chebyshev estimator.

There are two unimodal, positively skewed, one-parameter distributions with positive support. They are the Rayleigh and the half-normal distributions, both with one scale parameter σ. [For a monumental collection of continuous distributions, see Crooks (2014).] The former has a skewness of $g = 2\sqrt{\pi}(\pi - 3)/(4 - \pi)^{3/2} \simeq 0.631$, while for the observed one g is close to 1.15 (see Table 1). Also, the expected value μ and standard deviation σ are connected by the relation $\mu/\sigma = \sqrt{\pi/2} \simeq 1.571$, while for the observed distribution this ratio is close to 2.71. The latter has a skewness of $g = 2(4 - \pi)/(\pi - 2)^{3/2} \simeq 0.995$, quite close to the observed one. However, the mean-to-standard deviation ratio is $\mu/\sigma = \sqrt{2/\pi} \simeq 0.798$, very different from the observed one.

[2] Given a random variable X with cumulative distribution function $F(x)$, the test values x_α are defined by $F(x_\alpha) = 1 - \alpha$.

The only two-parameters candidates (with the characteristics before specified) are the (standard) gamma distribution with shape parameter α and scale parameter θ, and the (standard) beta prime distribution with shape parameters α and β. The skewness and the expected value-to-standard deviation ratio of these two distributions depend on both parameters, which can be then tuned to best approximate the observed distribution.

Proceeding as for the log-normal case, we find that, typically, the absolute value of the difference between the best-fit cumulative function and the observed one, $|\delta|_{\max}$, has maximum values of order of 10^{-2} for the gamma distribution. Accordingly, we did not investigate the gamma distribution case any further because the deviations from the observed distribution are much greater than the expected fluctuations due to the finite number of draws N.

For the case of the beta prime distribution, instead, the deviations are of order of $few \times 10^{-3}$. In the left panel of Figure 2, for example, we show the observed Cdf for $n = 500$ (black points), together with the best-fitted Cdf $F_{\beta'}(m_n)$ (continuous red line) of the beta prime distribution,

$$F_{\beta'}(m_n) = I\left(\frac{m_n}{1+m_n}, \alpha, \beta\right), \tag{4.1}$$

where $I(z,a,b) = B(z,a,b)/B(a,b)$ is the regularized incomplete beta function, $B(z,a,b) = \int_0^z dt\, t^{a-1}(1-t)^{b-1}$ is the incomplete beta function, and $B(a,b) = \int_0^1 dt\, t^{a-1}(1-t)^{b-1}$ is the Euler beta function (Abramowitz and Stegun, 1972). The values of the best-fit parameters are $\alpha = 12.77$ and $\beta = 23.33$, which give $|\delta|_{\max} = 0.0030$. For the beta prime case, then, we repeated the analysis done for the log-normal case. The minimum and maximum n considered are, again, 100 and 1000, respectively, and we proceeded with increments of $\Delta n = 10$. The best-fit values of the parameters α and β as a function of n are shown in the middle panel of Figure 2. The (red) continuous lines are the fit lines

$$\alpha = 12.54, \quad \beta = 22.78. \tag{4.2}$$

In the right panel of Figure 2, we show $|\delta|_{\max}$ as a function of n, The average value of $|\delta|_{\max}$, represented by the (red) continuous line, is about 0.0039. In Table 1, instead, we show the test values $m_{n,0.90}$, $m_{n,0.95}$, and $m_{n,0.99}$, the sample mean, the standard deviation, and the skewness of the Chebyshev distance statistic calculated by using the Cdf in Equation 4.1 with parameters in Equation 4.2. As it is clear from the table, the log-normal and beta prime distributions give compatible results. However, the log-normal distribution should be preferred to the beta prime because the former displays lower deviations from the observed distribution (lower values of $|\delta|_{\max}$) as is shown by the box-and-whisker plots in Figure 3.

5 Small n

One of the conditions of the emergence of Benford's law in data is that the numerical values are distributed across multiple orders of magnitude (Benford, 1938). Usually, but not always, this means that the law tends to be most accurate for numbers of data points relatively high ($n \geq 100$). However, "Benfordness" has been checked, and sometimes observed, also in small real-world data sets

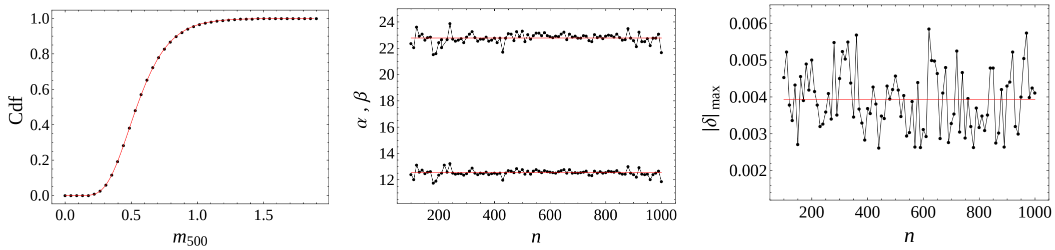

Figure 2: *Left panel.* The observed Cdf for $n = 500$ (black points), together with the best-fitted Cdf $F_{\beta'}(m_n)$ in Equation 4.1 (continuous red line). *Middle panel.* The best-fit values of the parameters α and β in Equation 4.1 as a function of n. The (red) continuous lines are the corresponding linear fits whose values are shown in Equation 4.2. *Right panel.* The maximum $|\delta|_{\max} = \max_{m_n} |\delta(m_n)|$ of the absolute value of the difference $\delta(m_n) = F_{\beta'}(m_n) - F_{\text{obs}}(m_n)$ between the theoretical Cdf with parameters in Equation 4.2 and the observed one as a function of n, together with its linear fit (the red continuous line).

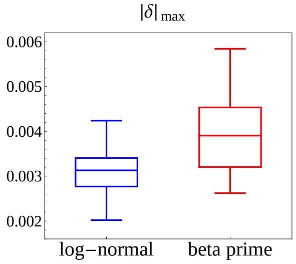

Figure 3: Box-and-whisker plots of the deviations $|\delta|_{\max}$ for both cases of log-normal and beta prime distributions.

($n < 100$), as for example in some of the financial data sets analyzed by Rodriguez (2004) (and re-analyzed by Lesperance et al., 2016).

For this reason, it is worth considering the case of “small n”. Our Monte Carlo simulation, as in the case of large n, is based on $N = 10^5$ draws for each sample size $10 \leq n \leq 99$. This time, we proceeded with size increments of $\Delta n = 1$. In the upper panels of Figure 4, we show the mean m_n, the standard deviation s_n, and the skewness g_n of the observed Cdf of the Chebyshev distance as a function of the sample size n. These estimators exhibit a regular dependence of the sample size n that can be quantified by the following best fit curves,

$$\overline{m_n} = 0.57 + 0.16 n^{-1}, \tag{5.1}$$

$$s_n = 0.21 + 0.13 n^{-1}, \tag{5.2}$$

$$g_n = 1.14 + 0.04 n^{-1}, \tag{5.3}$$

all represented in Figure 4 by (blue) continuous lines.

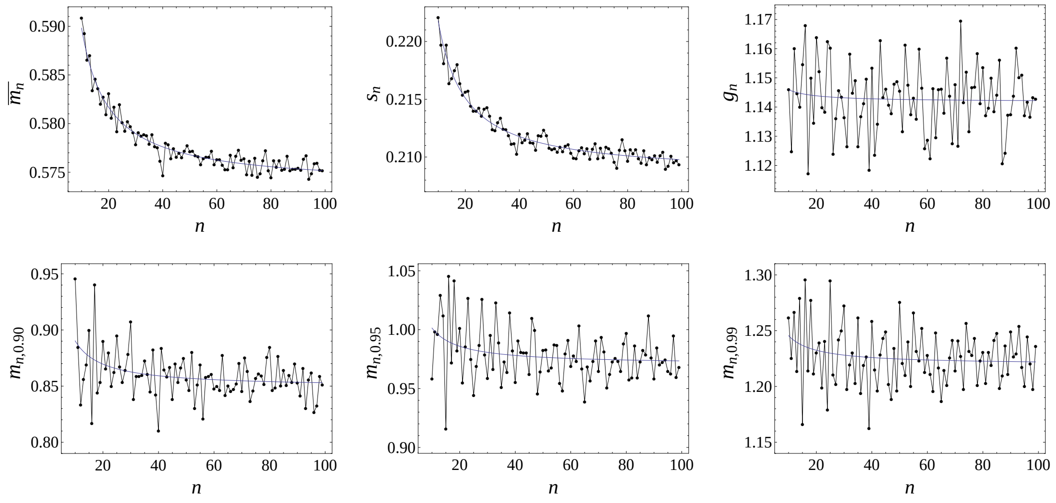

Figure 4: *Upper panels.* The mean m_n (left panel), the standard deviation s_n (middle panel), and the skewness g_n (right panel) of the observed Cdf of the Chebyshev distance as a function of the sample size $10 \leq n \leq 99$. *Lower panels.* As in the upper panels but for the test values $m_{n,90}$ (left panel), $m_{n,95}$ (middle panel), and $m_{n,95}$ (right panel).

In the lower panels of Figure 4, instead, we show our results for the test values $m_{n,90}$, $m_{n,95}$, and $m_{n,95}$. The (blue) lines are the fitting curves

$$m_{n,0.90} = 0.85 + 0.41n^{-1}, \tag{5.4}$$

$$m_{n,0.95} = 0.97 + 0.31n^{-1}, \tag{5.5}$$

$$m_{n,0.99} = 1.22 + 0.26n^{-1}. \tag{5.6}$$

Notice that the Cdf function of Chebyshev distance exhibits larger fluctuations (typically, of order of few percent) with respect to the asymptotic case $n \geq 100$. An analysis similar to that performed before, and aimed to find an empirical expression of the Cdf, is then not feasible in the small-n case. However, it is remarkable that the quantities in Equations 5.1-5.6 (with the exception of g_n and $m_{n,0.99}$) agree with the corresponding asymptotic values in Table 1 (log-normal case) for large values of n (approaching 99).

6 Efficacy: A Comparative Study

As already discussed in the Introduction, the most common tests in use for testing whether a numerical sample conforms to Benford's law are the Pearson's χ^2 (with 8 degrees of freedom), whose estimator is

$$\chi^2_{8,n} = n \sum_{d=1}^{9} \frac{[f_B(d) - f_n(d)]^2}{f_B(d)}\,, \tag{6.1}$$

and the Kolmogorov-Smirnov (Kolmogorov, 1933) test, based on the estimator

$$D_n = \sqrt{n} \max_d |F_B(d) - F_n(d)|, \tag{6.2}$$

where $F_B(d) = \log(1+d)$ and $F_n(d)$ are the Benford and the observed cumulative distribution functions, respectively.

The asymptotic 95%-confidence-level test value for the Pearson's χ^2 with 8 degrees of freedom is $\chi^2_{8,n,0.95} = 15.507$. The standard and Benford-specific asymptotic 95%-confidence-level test values for the Kolmogorov-Smirnov estimator are, instead, $D_{n,0.95} = 1.36$ (Smirnov, 1948) and $D^*_{n,0.95} = 1.15$ (Campanelli, 2024b), respectively.

In order to quantify the ability of a test to reject the null hypothesis H_0 (in our case Benford's law) given a set of data points with frequency distribution f_n, we introduce the quantity $n_{1-\alpha}$ as

$$n_{1-\alpha} = \{\text{number of data points needed to reject } H_0 \text{ at a significance level of } \alpha \,|\, f_n\}. \tag{6.3}$$

The larger is $n_{1-\alpha}$ the more conservative is the test in rejecting the null hypothesis given an observed distribution of data that deviates from the theoretical one. Alternatively, the smaller is $n_{1-\alpha}$ the greater is the efficacy of the test in rejecting the null.

The Benford distribution is a monotonically decreasing distribution, skewed to the write, and with a long tail. Since the median is $\sqrt{10} - 1 \simeq 2.1623$, we can define the body of the distribution as $d = 1, 2$, and the tail as $d = 3, 4, 5, 6, 7, 8, 9$. We expect the Pearson χ^2 test to be very sensitive to deviations in the tail of a Benford distribution because of the presence of the term $f_B(d)$ in the denominator of the expression of its estimator. On the other hand, because of the structure of its statistic, the Chebyshev test will be sensitive to variations in the body. Finally, the Kolmogorov-Smirvos test being based on the difference between the empirical and the Benford cumulative distribution functions, will be sensitive to cumulative (integrated) deviations in the observed frequency distribution from the theoretical one which in the following will be refereed to as “subductions”.

These expectations are confirmed by the results of a simulation where the Benford distribution is perturbed first in the body, then in the tail, and finally in an integrated way. The perturbed distributions are shown in Figure 5. Here, “IC” and “DC” stand for inverse and direct cascade, respectively. In an inverse cascade, part of the value of $f_B(d)$ for a particular digit is transferred to the previous digit, while in a direct cascade $f_B(d)$ is partially transferred to the next digit. In a subduction, a series of 2,3, or 4 consecutive values of $f_B(d)$ are partially transferred to the next ones (cumulative transfer of probabilities).

The perturbed distributions are reported in Table 2.

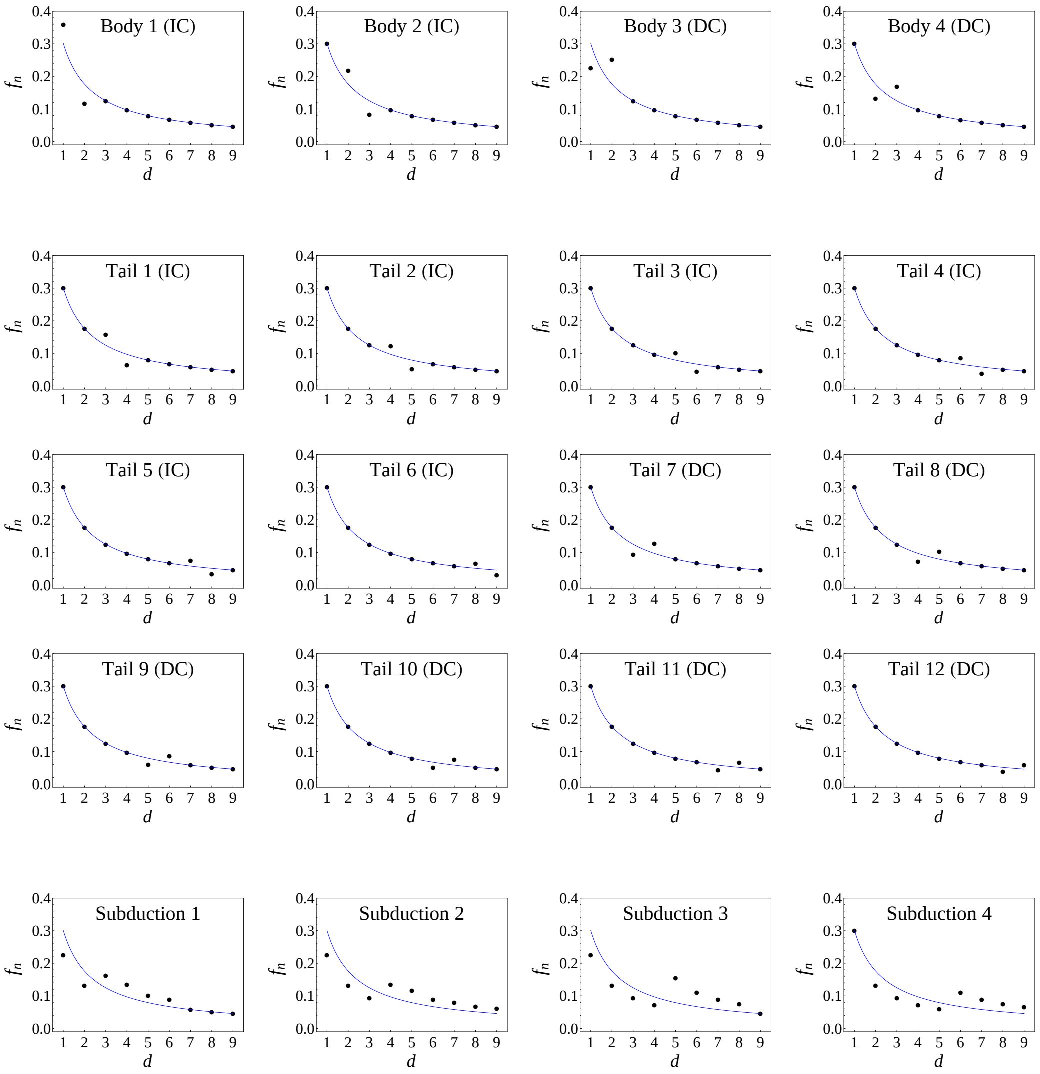

Figure 5: Perturbed Benford distributions. Perturbations are in the body ($d = 1, 2$; first row), in the tail ($d > 2$; second to fourth rows), and integrated ("subduction"; last row). "IC" and "DC" stand for inverse and direct cascade, respectively (see text for details).

Table 2: Simulated data in Figure 5: $f_n(d)$ and $f_d \equiv f_B(d)$ are the perturbed and standard Benford distributions, respectively.

Group	$f_n(1)$	$f_n(2)$	$f_n(3)$	$f_n(4)$	$f_n(5)$	$f_n(6)$	$f_n(7)$	$f_n(8)$	$f_n(9)$
Body									
1 (IC)	$f_1 + \frac{f_2}{3}$	$\frac{2}{3}f_2$	f_3	f_4	f_5	f_6	f_7	f_8	f_9
2 (IC)	f_1	$f_2 + \frac{f_3}{3}$	$\frac{2}{3}f_3$	f_4	f_5	f_6	f_7	f_8	f_9
3 (DC)	$\frac{3f_1}{4}$	$\frac{f_1}{4} + f_2$	f_3	f_4	f_5	f_6	f_7	f_8	f_9
4 (DC)	f_1	$\frac{3f_2}{4}$	$\frac{f_2}{4} + f_3$	f_4	f_5	f_6	f_7	f_8	f_9
Tail									
1 (IC)	f_1	f_2	$f_3 + \frac{f_4}{3}$	$\frac{2f_4}{3}$	f_5	f_6	f_7	f_8	f_9
2 (IC)	f_1	f_2	f_3	$f_4 + \frac{f_5}{3}$	$\frac{2f_5}{3}$	f_6	f_7	f_8	f_9
3 (IC)	f_1	f_2	f_3	f_4	$f_5 + \frac{f_6}{3}$	$\frac{2f_6}{3}$	f_7	f_8	f_9
4 (IC)	f_1	f_2	f_3	f_4	f_5	$f_6 + \frac{f_7}{3}$	$\frac{2f_7}{3}$	f_8	f_9
5 (IC)	f_1	f_2	f_3	f_4	f_5	f_6	$f_7 + \frac{f_8}{3}$	$\frac{2f_8}{3}$	f_9
6 (IC)	f_1	f_2	f_3	f_4	f_5	f_6	f_7	$f_8 + \frac{f_9}{3}$	$\frac{2f_9}{3}$
7 (DC)	f_1	f_2	$\frac{3f_3}{4}$	$\frac{f_3}{4} + f_4$	f_5	f_6	f_7	f_8	f_9
8 (DC)	f_1	f_2	f_3	$\frac{3f_4}{4}$	$\frac{f_4}{4} + f_5$	f_6	f_7	f_8	f_9
9 (DC)	f_1	f_2	f_3	f_4	$\frac{3f_5}{4}$	$\frac{f_5}{4} + f_6$	f_7	f_8	f_9
10 (DC)	f_1	f_2	f_3	f_4	f_5	$\frac{3f_6}{4}$	$\frac{f_6}{4} + f_7$	f_8	f_9
11 (DC)	f_1	f_2	f_3	f_4	f_5	f_6	$\frac{3f_7}{4}$	$\frac{f_7}{4} + f_8$	f_9
12 (DC)	f_1	f_2	f_3	f_4	f_5	f_6	f_7	$\frac{3f_8}{4}$	$\frac{f_8}{4} + f_9$
Subduction									
1	$f_1 - \frac{f_1}{4}$	$f_2 - \frac{f_2}{4}$	$f_3 + \frac{f_1}{8}$	$f_4 + \frac{f_1}{8}$	$f_5 + \frac{f_2}{8}$	$f_6 + \frac{f_2}{8}$	f_7	f_8	f_9
2	$f_1 - \frac{f_1}{4}$	$f_2 - \frac{f_2}{4}$	$f_3 - \frac{f_3}{4}$	$f_4 + \frac{f_1}{8}$	$f_5 + \frac{f_1}{8}$	$f_6 + \frac{f_2}{8}$	$f_7 + \frac{f_2}{8}$	$f_8 + \frac{f_3}{8}$	$f_9 + \frac{f_3}{8}$
3	$f_1 - \frac{f_1}{4}$	$f_2 - \frac{f_2}{4}$	$f_3 - \frac{f_3}{4}$	$f_4 - \frac{f_4}{4}$	$f_5 + \frac{f_1}{4}$	$f_6 + \frac{f_2}{4}$	$f_7 + \frac{f_3}{4}$	$f_8 + \frac{f_4}{4}$	f_9
4	f_1	$f_2 - \frac{f_2}{4}$	$f_3 - \frac{f_3}{4}$	$f_4 - \frac{f_4}{4}$	$f_5 - \frac{f_5}{4}$	$f_6 + \frac{f_2}{4}$	$f_7 + \frac{f_3}{4}$	$f_8 + \frac{f_4}{4}$	$f_9 + \frac{f_5}{4}$

Table 3: Simulated data in Figure 5: number of data points $n_{0.95}$ needed to reject the null hypothesis (Benford's law) at 95% confidence level by a given statistical test. The tests are: Chebyshev (m_n), Pearson χ^2 ($\chi^2_{8,n}$), and Kolmogorov-Smirnov (D_n for the standard continuous case, and D^*_n for the Benford-specific one.)

Group	m_n	$\chi^2_{8,n}$	D_n	D^*_n
Body				
1 (IC)	274	501	537	384
2 (IC)	543	654	1067	763
3 (DC)	167	305	327	234
4 (DC)	486	585	955	683
Tail				
1 (IC)	902	812	1773	1268
2 (IC)	1351	971	2656	1899
3 (IC)	1890	1130	3715	2656
4 (IC)	2518	1290	4950	3540
5 (IC)	3237	1450	6362	4549
6 (IC)	4045	1610	7951	5685
7 (DC)	965	868	1896	1356
8 (DC)	1603	1152	3152	2254
9 (DC)	2402	1436	4721	3375
10 (DC)	3359	1721	6603	4722
11 (DC)	4477	2006	8800	6292
12 (DC)	5754	2291	11311	8087
Subduction				
1	167	225	130	93
2	167	162	82	59
3	167	90	61	44
4	167	163	130	93

In Table 3, instead, we show the number of data points $n_{0.95}$ needed to reject the null hypothesis (Benford's law) by a given statistical test at 95% confidence level. Notice that the name of the test in Table 3 is the same as the symbol of the corresponding statistical estimator. Moreover, D_n and D^*_n stand for the classical and Benford-specific Kolmogorov-Smirnov tests, respectively.

As it is clear from Table 3, the Benford-specific Kolmogorov-Smirnov test is always more sensitive than the classical one and, at the same time, the most effective in detecting subductions. Moreover, the Chebyshev and the Pearson χ^2 tests are, in a certain way, complementary in the sense that their efficacy in rejecting Benford's law in data that deviate somehow from the theoretical expectation depends on the position, in the body or tail of the distribution, of such deviations.

7 Power: A Comparative Study

We now consider the power of the Chebyshev test when testing Benford's law and compare it to the power of the Pearson's χ^2 and Kolmogorov-Smirnov tests (the standard and the Benford-specific ones). Given a null hypothesis H_0 (in our case Benford's law), the power of a test against an alternative hypothesis H_1 at a given significance level of α is

$$B_{1-\alpha} = \Pr(\text{reject } H_0 \text{ at a significance level of } \alpha \mid H_1 \text{ is true}). \tag{7.1}$$

The "natural" alternative hypothesis to Benford's law is the so-called "generalized Benford's law". This is for the following reason. Benford's law on the first-digit distribution of a numerical data set emerges if the underlying distribution of the numerical data is scale and base invariant. If the distribution is only scale invariant, then the first-digit distribution of the numerical values will follow a generalized Benford's law. It is reasonable to assume that numerical data coming from "natural phenomena" and which do not depend on a particular scale (like lengths of rivers, fundamental constants, etc.), are also base invariant, thus producing Benford's law. On the other hand, "human phenomena" (like those connected to political elections, campaign finance, etc.) even if scale-invariant are not *necessarily* base invariant, thus producing a generalized Benford's law.

Let us consider a probability distribution function $f(x)$ with support $[a, b]$, where $a, b \in \mathbb{R}^*$. Let us assume that $f(x)$ is a homogeneous function of degree $\gamma - 1$, where $\gamma \in \mathbb{R}^*$ (namely a scale-invariant function). Then the first-digit distribution of the values of $f(x)$, $f_{GB}(d)$, follows a generalized Benford's law:

$$f_{GB}(d) = \frac{(d+1)^\gamma - d^\gamma}{10^\gamma - 1}. \tag{7.2}$$

Notice that for $\gamma = 1$, the probability mass function is uniform, $f_{GB}(d) = 1/9$, while for $\gamma \to 0$, the distribution $f_{GB}(d)$ approaches Benford's law in Equation 2.2.

In the left panel of Figure 6, we show the probability mass function of a generalized Benford variable for different values of the parameter γ (from top to bottom: $\gamma = -1, -0.9, ..., 0.9, 1$) together with Benford's law (the blue continuous line).

In the middle panel of Figure 6, we show the simulated power obtained from a Monte Carlo simulation with $N = 10^5$ random samples, each of size $n = 250$, extracted from a generalized Benford distribution for different values of γ. In the right panel, we show the results of a similar simulation, this time with $n = 50$. As it is clear from the figure, the power of the Benford-specific Kolmogorov-Smirnov test is always greater than the corresponding power for the classical Kolmogorov-Smirnov test. Moreover, the former has the highest power among the tests here considered. This is explained by the fact that a generalized Benford distribution behaves like a "subduction" of a standard Benford

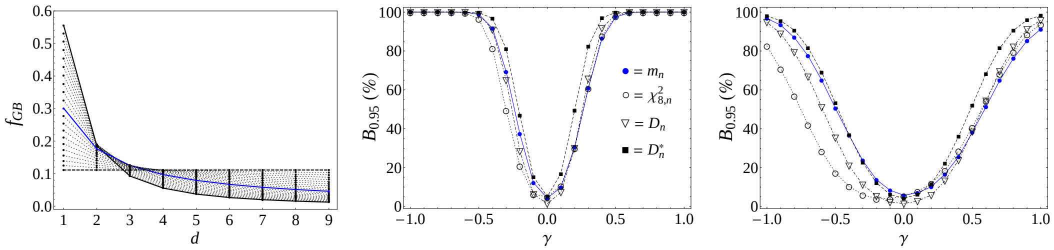

Figure 6: *Left panel.* Generalized Benford distribution for different values of γ. From top to bottom: $\gamma = -1, -0.9, ..., 0.9, 1$. The (blue) continuous line is Benford's law, and the horizontal dashed line is the discrete uniform distribution (corresponding to $\gamma = 1$). *Middle panel.* Simulated power at a significance level of $\alpha = 0.05$ for a sample size of $n = 250$ for four different statistical tests: Chebyshev test (m_n), Pearson's χ^2 test ($\chi^2_{8,n}$), standard Kolmogorov-Smirnov test (D_n), and Benford-specific Kolmogorov-Smirnov test (D^*_n). The null hypothesis is Benford's law, while the alternative hypothesis is the generalized Benford's law for different values of the parameter γ. *Right panel.* As in the middle panel for $n = 50$.

distribution, and then the Kolmogorov-Smirnov test is very effective in rejecting the null (Benford's law) if the alternative hypothesis (a generalized Benford's law) is true.

For low values of n the power of the χ^2 test is low when $\gamma < 0$ compared to the Chebyshev test. This is because a generalized Benford distribution with negative parameter γ exhibits relatively small deviations in the tail and relatively large deviations in the body when compared to a (standard) Benford distribution. In this case, then, the Chebyshev test is more effective than the χ^2 test in rejecting Benford's law if the alternative hypothesis of a generalized Benford's law is true.

8 Application: Annual Deaths Counts by Country

As an application of the Chebyshev distance to the goodness-of-fit for the Benford distribution, we consider the following data sets: i) the annual all-causes-deaths counts by country (World Population Prospects, 2024), ii) the annual number of homicides by country (United Nations Office on Drugs and Crime, 2024), iii) the annual number of deaths from infectious diseases by country (IHME, Global Burden of Disease, 2024a), and iv) the annual suicide counts by country (IHME, Global Burden of Disease, 2024b).

In order to study the first-digit distribution of the above data sets, we use the Chebyshev, Pearson χ^2, and Kolmogorov-Smirnov statistics. The evaluation of the p values for the three statistics is based on the corresponding cumulative distribution functions, $p = 1 - \text{Cdf}(X_n)$, where $X_n = m_n, \chi^2_{8,n}, D_n$. For the Chebyshev case, we use Equation 2.3 with parameters in Equation 3.1, while for the Pearson χ^2 with 8 degrees of freedom we use the standard asymptotic Cdf

$$\phi(\chi^2_{8,n}) = 1 - Q(4, \chi^2_{8,n}), \tag{8.1}$$

where $Q(a,z) = \Gamma(a,z)/\Gamma(a)$ is the regularized incomplete gamma function, $\Gamma(a,z) = \int_z^\infty dt\, t^{a-1} e^{-t}$ is the incomplete gamma function, and $\Gamma(a) = \Gamma(a,0)$ is the (standard) gamma function (Abramowitz and Stegun, 1972).

The asymptotic Cdf (Kolmogorov, 1933) of the Kolmogorov-Smirnov statistic D_n, $\Phi(D_n)$, can be written in term of the theta function as (Campanelli, 2024b)

$$\Phi(D_n) = \vartheta_4\Big(0, e^{-2D_n^2}\Big), \tag{8.2}$$

where $\vartheta_4(z,q) = 1 + 2\sum_{k=1}^{+\infty}(-1)^k q^{k^2} \cos(2kz)$ is the theta function of type 4, argument z, and nome q (Abramowitz and Stegun, 1972). Equation 8.2 is generally used for $n \geq 35$ (Smirnov, 1948) and was obtained in the hypothesis of continuous random variables. For the discrete Benford random variable, the asymptotic Kolmogorov Cdf shows unacceptable large deviations, up to about 35%, from the ones observed in Monte Carlo simulations (Campanelli, 2024b). Such deviations can be reduced to a level below 0.5$\%$ if the following linear transformation of the argument of the Kolmogorov Cdf is performed (Campanelli, 2024b),

$$\Phi(D_n) \to \Phi^*(D_n) = \Phi(aD_n + b), \tag{8.3}$$

with $a = 0.984$, and $b = 0.227$.

In Figure 7, we show the observed first-digit frequency distributions of annual case counts by country for the four data sets and for four selected years (2021, 2020, 2019, and 2018) superimposed to Benford's law (the blue continuous line).

All data sets comply with Benford' law, as it is clear from Table 4, where we show the p values obtained from the three statistical tests together with the number of data points n (equal to the number of countries) and the range of the numerical data $[\min, \max]$. The only exception to such a compliance is represented by the deaths from infectious diseases in 2020. While the Chebyshev and Pearson χ^2 tests cannot reject the null hypothesis of Benford's law at a significance level of $\alpha = 0.05$, the Kolmogorov-Smirnov test gives a p value below 0.01. Looking at Figure 7, it is clear that this is a typical "subduction" case: the values of the observed frequencies for the first four digits are well below the theoretical expectations and, at the same time, there is an excess in frequency for the last five digits. This explains why the Kolmogorov-Smirnov is more effective in rejecting the null than the other two tests (the numbers of data points needed to achieve such a low p value would be $n = 1315$ and $n = 332$ for the Chebyshev and Pearson χ^2 tests, respectively.)

For the case of the deaths from infectious diseases in 2018 and 2019, large deviations from Benford's law are observed in the tail of the distributions (see Figure 7). This explain why the χ^2 test is the least conservative and give the lowest p values among the three tests. On the other hand, the distributions of homicides in 2021 and deaths from infectious diseases in 2021 exhibit relatively large deviations in the body ($d = 1, 2$). In this case, and as to be expected, the less conservative test is the Chebyshev one.

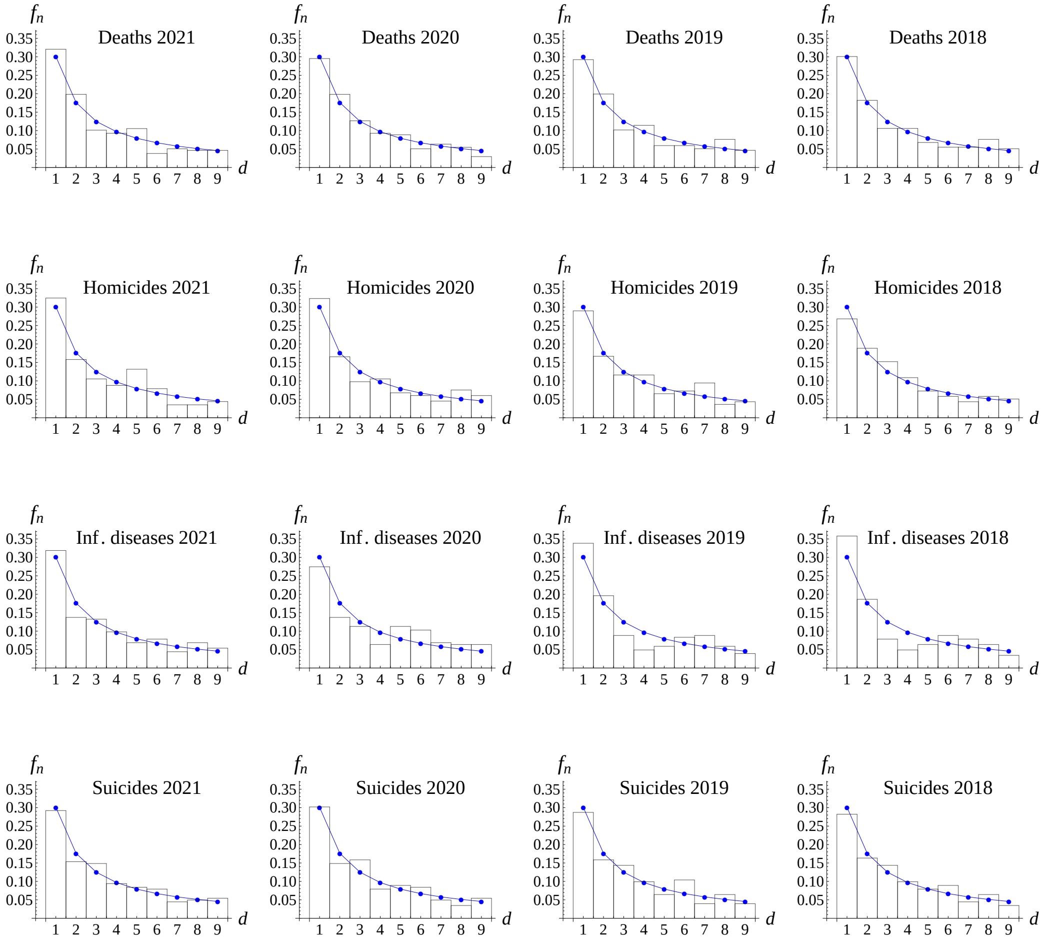

Figure 7: Observed first-digit frequencies of the annual deaths by country. First row: all-causes-deaths. Second row: homicides. Third row: deaths from infectious diseases. Fourth row: suicides. The (blue) continuous lines represent Benford's law.

9 Discussion and Conclusions

The origin of the emergence of Benford's law in real-data sources is still an open problem in statistics. Nevertheless, this law has been widely used to flag anomalies in numerical data distributions coming from very disparate disciplines, such as epidemiology, cryptology, finance, politics, and data imaging, to cite a few.

Standard test statistics, such as the Kolmogorov-Smirnov and the Pearson χ^2 statistics, are routinely used to test the compliance to Benford's law in data. However, the Kolmogorov-Smirnov test

Table 4: p values of the observed first-digit frequencies of the annual deaths by country for different statistics: the Chebyshev (m_n), the Pearson χ^2 ($\chi^2_{8,n}$), and Kolmogorov-Smirnov (D_n for the standard continuous case, and D^*_n for the Benford-specific one.) Also indicated is the range of deaths, $[\min, \max]$, and the number of data points n (equal to the number of countries).

Year	n	$[\min, \max]$	m_n	$\chi^2_{8,n}$	D_n	D^*_n
All causes deaths						
2021	237	[2,13093783]	0.70	0.49	0.80	0.45
2020	237	[1,10411387]	0.90	0.90	1.00	0.89
2019	236	[3,10128803]	0.83	0.53	1.00	0.84
2018	236	[2,9967030]	0.83	0.77	0.98	0.74
Homicides						
2021	114	[1,241467]	0.46	0.57	0.99	0.78
2020	133	[2,250505]	0.94	0.86	0.99	0.77
2019	138	[2,251483]	0.75	0.77	1.00	0.90
2018	138	[2,303601]	0.82	0.96	1.00	0.85
Deaths by infectious diseases						
2021	204	[1,3261677]	0.47	0.76	1.00	0.88
2020	204	[1,2586834]	0.47	0.079	0.013	0.0035
2019	204	[1,1751094]	0.25	0.083	0.52	0.24
2018	204	[1,1816219]	0.13	0.053	0.32	0.13
Suicides						
2021	202	[1,188578]	0.91	0.93	0.99	0.76
2020	202	[1,186195]	0.64	0.63	1.00	0.87
2019	202	[1,183296]	0.53	0.45	0.99	0.76
2018	202	[1,186055]	0.93	0.83	0.99	0.76

is too conservative for testing discrete distributions as Benford's, while the Pearson χ^2 test has a high power only for large sample sizes.

The former problem has been recently solved by the author (Campanelli, 2022d) by showing that an appropriate linear transformation of the argument of the Kolmogorov cumulative distribution function makes the Kolmogorov-Smirnov test accurate at a level of 1% when testing Benford's law for moderately large and large numbers of data points. On the other hand, other authors have tried to overcome the limitations of standard statistical tests by considering new statistics, such as the Chebyshev distance statistic. Introduced by Leemis et al. (2000), the properties of this new estimator have been subsequently studied by Morrow (2014), who has provided the corresponding asymptotic test values.

In this paper, we have extended the work by Morrow by finding, by means of a Monte Carlo simulation, an empirical expression of the asymptotic cumulative distribution function of the Chebyshev distance. Our results show that the statistical test based on the Chebyshev distance statistic is accurate at a level of 1% when testing Benford's law for moderately large and large numbers of data points, $n \geq 100$.

For small values of n, the Chebyshev distance exhibits larger fluctuations (of order of few percent) with respect to the asymptotic case. For this reason, the search for a (n-dependent) empirical expression of the Cdf was not pursued and only test values of the Chebyshev distance as a function of the sample size were estimated empirically by performing a Monte Carlo simulation.

The performance (efficacy and power) of the goodness-of-fit test based on the Chebyshev statistic was analyzed and compared to that of the Pearson χ^2, and to that of both the classical Kolmogorov-Smirnov test and the Benford-specific one. The general result is that the Benford-specific Kolmogorov-Smirnov test is always more effective and powerful than the classical one and then should be preferred over the latter when testing Benford's law. Moreover, the χ^2 and Chebyshev tests are "complementary": the former performs well when the deviations of the observed distribution from the Benford one are primarily in the tail (digits $d \geq 3$), while if they are in the body ($d = 1, 2$), the Chebyshev test outperforms the χ^2 test. Finally, the (Benford-specific) Kolmogorov-Smirnov test is the most effective in detecting cumulative (integrated) deviations in the observed frequency distribution from the theoretical one.

As an application of the Chebyshev test to Benford's law, we considered the first-digit distributions of the annual deaths counts by country (all-causes, homicides, from infectious diseases, and suicides) in four different years. All these distributions comply with Benford's law to a very high level of confidence. As for the case of efficacy and power, however, the Chebyshev, χ^2, and Benford-specific Kolmogorov-Smirnov tests perform (slightly) differently depending on the particular observed distribution: a combined use of these three tests is then highly recommended when checking for conformance to Benford's law.

References

Abramowitz, M., Stegun, I. A. (eds.), (1972), "Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables," New York: Dover Publications.

Benford, F., (1938), "The Law of Anomalous Numbers," *Proceedings of the American Physical Society*, 78: 551-572.

Campanelli, L., (2022a), "A Statistical Cryptanalysis of the Beale Ciphers," *Cryptologia*, 47(5), 466-473.

Campanelli, L. (2022b), "On the Euclidean Distance Statistic of Benford's Law," *Communications in Statistics - Theory and Methods*, 53(2), 451-474.

Campanelli, L., (2022c). “Testing Benford’s Law: from small to very large data sets,” *Spanish Journal of Statistics*, 4: 41-54.

Campanelli, L. (2023), “Breaking Benford’s law: A statistical analysis of Covid-19 data using the Euclidean distance statistic,” *Statistics in Transition new series*, 24(2), 201-215.

Campanelli, L. (2024a), “Monkeypox Obeys the (Benford’s) Law: A Dynamic Analysis of Daily Case Counts in the United States of America,” *Statistics in Transition new series*, 25(2), 219-227.

Campanelli, L. (2024b), “Tuning up the Kolmogorov-Smirnov test for testing Benford’s law,” *Communications in Statistics - Theory and Methods*, 1-0. doi:10.1080/03610926.2024.2318608.

Cho, W. K. T., Gaines, B. J., (2007), “Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance,” *American Statistician*, 61: 218-223.

Crooks, G. E., (2014), “Field Guide to Continuous Probability Distributions,” ISBN: 978-1-7339381-0-5. https://threeplusone.com/fieldguide.

Farhadi, N., (2021), “Can we rely on COVID-19 data? An assessment od data from over 200 countries worldwide,” *Science Progress*, 104: 1-19.

Leemis, L. M., Schmeiser, B. W., Evans, D. L., (2000), “Survival Distributions Satisfying Benford’s Law,” *American Statistician* 54: 236-2,41.

Lesperance M., Reed W. J., Stephens M. A., Tsa, C., Wilton B., (2016), “Assessing Conformance with Benford’s Law: Goodness-Of-Fit Tests and Simultaneous Confidence Intervals,” *PLos ONE*, 11(3): e0151235. DOI: 10.1371/journal.pone.0151235.

Miller, S. J. (ed.), 2015, “Benford’s Law: Theory and Applications,” Princeton. Princeton University Press.

Morrow, J., (2014), “Benford’s Law, Families of Distributions and a Test Basis,” London: Centre for Economic Performance.

Noether, G. E., (1963), “Note on the Kolmogorov statistic in the discrete case,” *Metrika*, 7: 115-116.

Rodriguez, R. J., (2004), “Reducing false alarms in the detection of human influence on data,” *Journal of Accounting, Auditing & Finance*, 19: 141-158.

Roukema, B. F., (2013), “A first-digit anomaly in the 2009 Iranian presidential election,” *Journal of Applied Statistics*, 41: 1, 164-199.

Sambridge, M., Jackson, A., (2020), “National COVID numbers - Benford’s law looks for errors,” *Nature*, 581: 384.

Smirnov, N., (1948), “Table for estimating the goodness of fit of empirical distributions,” *Annals of Mathematical Statistics*, 19, 279-281.

World Population Prospects, (2024), Processed by Our World in Data. “Number of deaths, medium projection - UN WPP”. United Nations. “World Population Prospects [original data]. Retrieved December 27, 2024 from https://ourworldindata.org/grapher/number-of-deaths-per-year

United Nations Office on Drugs and Crime, (2024), Population based on various sources (2023) - with major processing by Our World in Data. “Number of homicides -sex: Total - age: Total” [dataset]. United Nations Office on Drugs and Crime, “United Nations Office on Drugs and Crime - Intentional Homicide Victims”; Various sources, “Population” [original data]. Retrieved December 28, 2024 from `https://ourworldindata.org/homicides`.

IHME, Global Burden of Disease, (2024a), Processed by Our World in Data. “Deaths from infectious diseases” [dataset]. IHME, Global Burden of Disease, “Global Burden of Disease - Deaths and DALYs” [original data]. Retrieved December 29, 2024 from `https://ourworldindata.org/grapher/deaths-from-infectious-diseases`.

IHME, Global Burden of Disease, (2024b), Processed by Our World in Data. “Total umber of deaths from self-harm” [dataset]. IHME, Global Burden of Disease, “Global Burden of Disease - Deaths and DALYs” [original data]. Retrieved January 2, 2025 from `https://ourworldindata.org/suicide`.

Wase V., (2021), “Benford's law in the Beale ciphers,”, *Cryptologia* 45: 3, 282-286.

Wolfram Research, Inc., (2023), *Mathematica*, Version 7.0, Champaign, IL.

Received: August 30, 2024

Accepted: February 16, 2025