STATISTICS IN TRANSITION new series, June 2024
Vol. 24, No. X, pp. 1–XX, DOI 10.59170/stattrans-2023-xxx
Received – xx.xx.xxxx; accepted – xx.xx.xxxx

Monkeypox obeys the (Benford) law: a dynamic analysis of daily case counts in the United States of America

Leonardo Campanelli[1]

ABSTRACT

We analyze, for the first time, the first-digit distribution of the monkeypox daily cases in the United States of America, from May 17 to September 21, 2022. The overall data follow Benford's law, a conclusion substantiated by eight different statistical tests, including the "Euclidean distance test", which has been designed to specifically check Benford's distribution in data. This result aligns with those of other infectious diseases, such as COVID 19, whose Benfordness has already been confirmed in the literature. Daily counts of monkeypox cases, like any other disease evolve in time. For this reason, we analyzed the temporal deviation of monkeypox counts from Benford's law to check for possible anomalies in the temporal series of cases. The dynamic analysis was performed by means of the Euclidean distance test. This is because, to our best knowledge, that is the only statistically valid, Benford-specific test whose underlying estimator has a cumulative distribution function with known analytical properties, and is applicable to small and large samples. This is the case in dynamic analyses, where the number of data points usually starts from small values and then increases in time. No anomalies were detected, which indicates that no (fraudulent) alterations or errors in data gathering took place.

Key words: Benford's law, monkeypox, Euclidean distance statistic, dynamic analysis

1. Introduction

Monkeypox is a viral zoonotic infectious disease caused by a virus in the genus Orthopoxvirus. An ongoing outbreak started on May 6, 2022 in London, United Kingdom. From May 18 onwards, cases were reported worldwide in more than about 100 countries. This is the first time monkeypox has spread outside Central (Congo Basin Clade) and West Africa (West African Clade), where the disease is endemic (WHO, 2022).

There is evidence that the spread of infectious diseases conforms to Benford's law. Indeed, Sambridge et al. (2010) found that the total numbers of cases of 18 infectious diseases reported to the World Health Organization (WHO) by 193 countries worldwide in 2007 follow a Benford's distribution. Recently, Benford's law has been applied to the study of COVID-19 data, in particular to daily, weekly, and cumulative case and death counts of various countries [see, e.g., Sambridge and Jackson (2020), Farhadi (2021), and Campanelli (2022a).] The general result is that the Benford's distribution well describes the first-digit distributions of COVID-19 data for most of the countries and, then, it can be used to flag "anomalies" in the data of specific countries.

[1] All Saints University School of Medicine, Canada. E-mail: leonardo.s.campanelli@gmail.com. ORCID: https://orcid.org/0000-0002-7200-9990.

Benford's law (Benford, 1938) is an empirical statistical law according to which the probability $P_B(d)$ of occurrence of the first significant digit d in "particular" data sets is

$$P_B(d) = \log\left(1 + \frac{1}{d}\right). \tag{1}$$

Although it is now known that some distributions satisfy Benford's law [see, e.g. Morrow (2014) and references therein] and that particular principles lead to the emergence of the Benford's phenomenon in data (Hill, 1995a, 1995b, and 1995c), no general criteria has be found that fully explain when and why Benford's law holds for a "generic" set of data.

Although much work is still needed to understand the theoretical basis of the law, the number of its applications has grown in the last few decades [for theoretical insights and general applications of Benford's law see, e.g. Miller (2015)]. Probably, the most famous applications are to detecting tax (see e.g. Nigrini, 1996), campaign finance (see e.g., Cho and Gaines, 2007), and election (see e.g. Roukema, 2013) frauds. Other interesting applications are in image processing (Pérez-González et al., 2007), where Benford's law can be used to test whether or not the image has been compressed, in natural sciences, where the law has been shown to hold for geophysical observables such as the depths of earthquakes (Sambridge et al., 2010), and in cryptology, where it can be used to examine the truthfulness of undeciphered numerical codes (Wase, 2021, Campanelli, 2022b). [2]

The aim of this paper is to assert if the data relative to the monkeypox daily counts in the United States of America (USA) comply or not with Benford's law. The motivation behind this is that, as already noticed, there are already sufficient indications that the number counts of (confirmed and/or death) cases for other infectious diseases (notably COVID 19) follow a Benford distribution. Therefore, a departure of monkeypox data from Benford's law could signal an anomaly, and eventually a fraud, in the data.

2. Method

It is well known that the compliance of data sets to Benford's law improves as the range of the data increases. Daily and cumulative death cases by country are then not appropriate when checking for the compliance of the monkeypox first-digit distributions to Benford's law because there have been only few tens of deaths worldwide since the start of the outbreak (WHO, 2022). Another possibility would be the use of cumulative confirmed case counts. The disadvantage of using this type of data is that as cumulative case numbers begin to flatten (e.g. after a monkeypox "wave" has passed), first digits tend to become all the same, thus distorting relative digit frequencies. In order to overcome this problem, we will only analyze the data on daily confirmed cases by country. However, the only country with daily case numbers which extend on a statistically appreciable range is the USA: Here, the data cover about three orders of magnitude, while in all the other countries affected

[2] The number of articles, books, and other resources related to Benford's law is enormous. Only in 2021 (the year preceding the writing of the present article), for example, the total number of articles, preprints, proceedings, research reports, books chapters, bachelor, master and PhD theses directly connected to Benford's law was (greater than) 131 (Berger et al., 2009). The interested reader can refer the "Benford online bibliography" (Berger et al., 2009) for an up-to-date collection of Benford's-law-related works.

by monkeypox they extend at most on two (WHO, 2022). Accordingly, we will focus our analysis on the daily case counts from the USA.

3. Results

In order to check the conformance of monkeypox data to Benford's law, we will use the "Euclidean distance test", which has been recently proposed by the author to specifically quantify the goodness of fit of a data sample to Benford's law (Campanelli, 2022c). The reasons behind this choice are discussed in Section 4.

3.1. Overall analysis

The Euclidean distance test is based on the Euclidean distance estimator d_N^*, first introduced by Cho and Gaines (2007) and then analyzed by Morrow (2014),

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

where $P(d)$ is the observed first-digit frequency distribution of a sample of size N. The (empirical) cumulative distribution function (CDF) of the Euclidean distance statistic found by the author (Campanelli, 2022c) allow us to evaluate p values as $p = 1 - \text{CDF}(d_N^*)$.

Data of the 2022 USA monkeypox outbreak are from the Centers for Disease Control and Prevention (CDC, 2022) and are updated to September 21, 2022. They are the confirmed daily cases reported to the CDC since May 17, 2022, the start of the response to the current outbreak. They include either the positive laboratory test report date, CDC call center reporting date, or the case data entry date into CDC's emergency response common operating platform.

In Table 1, we show the range of daily cases, $[\min, \max]$, the number of days, N, the Euclidean distance, d_N^*, and the corresponding p value. In the left panel of Figure 1, instead, we show the observed first-digit frequency distribution of daily case counts superimposed to Benford's law. As it is clear from the table and figure, the data comply with Benford's law at a high level of significance.

Table 1. The Euclidean distance d_N^* in Equation (2) and its corresponding p value for the first-digit distribution of the monkeypox daily case counts in the USA. Also indicated are the range of cases, $[\min, \max]$, and the number of days, N. Counts are from the CDC (2022) and are updated to September 21, 2022. The last three columns show the reduced χ^2 score, $\chi^2_{\text{red}} = \chi^2/\nu$, the number ν of degrees of freedom, and the p value, $p(\chi^2)$, of the χ^2 statistic defined in Equation (4).

Range	N	d_N^*	p	χ^2_{red}	ν	$p(\chi^2)$
$[1, 916]$	125	1.0031	0.284	0.5462	76	0.9996

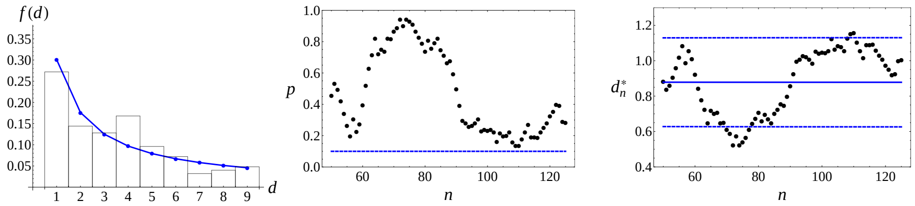

Figure 1: *Left panel.* Observed first-digit frequencies of the monkeypox daily case counts in the USA. The (blue) continuous line represents Benford's law. *Middle panel.* p values of the Euclidean distance statistic d_n^* as a function of the number of data points n (number of days). The (blue) dashed line is $p = 0.10$. *Right panel.* The Euclidean distance statistic as a function of n. The (blue) continuous line represents the expected value of d_n^* for a Benford's distribution, while the (blue) dashed lines show the corresponding one-sigma interval.

3.2. Dynamic analysis

Since the daily counts relative to monkeypox, as well as to any other infectious disease, evolve in time, it is interesting and statistically befitting to quantify the deviation of the timeline of those counts from Benford's law. Indeed, a dynamic data analysis of the chronology of the counts better captures the statistical properties of the spread of a disease.

Such a dynamic analysis can be performed by considering the following χ^2 statistic:

$$\chi^2 = \sum_{n=N'}^{N} \left(\frac{d_n^* - \overline{d_n^*}}{\sigma_n} \right)^2 . \tag{3}$$

Here, $\overline{d_n^*}$ and σ_n are the expected value and standard deviation of the Euclidean distance statistic for the Benford's distribution (Campanelli, 2022c), while d_n^* is the value of the observed Euclidean distance statistic for n data points (the ordinal number of days in our case).

As already noticed, the compliance to Benford's law improves as the range of the data increases. For this reason, we let the sum in Equation (4) to begin from N', the day starting from which the data range extends at least on two orders of magnitude (in the case at hand, $N' = 50$). The number ν of degrees of freedom for the χ^2 statistic is then $\nu = N - N'$.

In the middle and right panels of Figure 1 we show, respectively, the p values and scores of the Euclidean distance statistic d_n^* as a function of n. The (blue) continuous line in the right panel represents the expected value of d_n^* for a Benford's distribution, while the (blue) dashed lines show the corresponding one-sigma interval.

As it is clear from the figure, the null hypothesis of conformance to Benford's law can never be rejected at a 10% level of significance. Moreover, the values of the observed d_n^* are relatively close to the ones expected for a Benford's distribution. This closeness can be quantified by using Equation (3), which gives $\chi^2 = 41.51$ for 76 degrees of freedom. This corresponds to a reduced χ^2 score as low as $\chi^2_{\text{red}} = \chi^2/\nu = 0.5462$ and to a p value as large as $p(\chi^2) = 0.9996$. These values for the reduced χ^2 and p values, reported in Table 1 for convenience, show that the temporal series of the monkeypox daily case counts in the USA conforms to the Benford's distribution to a very high significance level.

4. Discussion

The most common test in use for testing whether a numerical sample satisfies Benford's law is the Pearson's χ^2 (with 8 degrees of freedom), whose estimator is

$$\chi_8^2 = N \sum_{d=1}^{9} \frac{[P(d) - P_B(d)]^2}{P_B(d)}. \tag{4}$$

The well-known problem with this statistic is the low power at small N ($N < 100$) (see, e.g., Morrow, 2014).

To overcome this problem, other more powerful test statistics like the Kolmogorov-Smirnov (Kolmogorov, 1933) and Kuiper (Kuiper, 1960) statistics have been used. However, these statistics have been constructed for continuous distributions and are generally conservative when testing discrete distributions (see, e.g., Morrow, 2014). Only recently, Benford-specific asymptotic test values have been found by Morrow (2014). The Kolmogorov-Smirnov and Kuiper statistics, D_N^* and V_N^* respectively, are defined as follows. Let D_N^+ and D_N^- be $D_N^\pm = \max_d[\pm Z(d)]$, where $Z(d) = S(d) - T(d)$, $S(d) = \sum_{d'=1}^{d} P(d')$, and $T(d) = \sum_{d'=1}^{d} P_B(d')$. Then,

$$D_N^* = (\sqrt{N} + \delta_N) \max[D_N^+, D_N^-] \tag{5}$$

and

$$V_N^* = (\sqrt{N} + \nu_N)(D_N^+ + D_N^-). \tag{6}$$

Here, the correction terms $\delta_N = 0.12 + 0.11/\sqrt{N}$ and $\nu_N = 0.155 + 0.24/\sqrt{N}$ have been introduced by Stephens (1970) for the case of continuous distributions to produce accurate test statistics regardless of the sample size. Asymptotic test values for these statistics are given in Table 2. However, since both the asymptotic and the small-sample form of the CDF for the case of a Benford's distribution are still unknown, p values cannot be calculated. A dynamic analysis of monkeypox cases is then not feasible in this case.

Another statistic used to test Benford's law is the so-called "max statistic". It is defined by

$$m_N^* = \sqrt{N} \max_d [P(d) - P_B(d)]. \tag{7}$$

Introduced by Leemis et al. (2000) to specifically test Benford' law, the statistical properties of the max estimator were subsequently analyzed by Morrow (2014), who provided asymptotic test values (reported in Table 2). As for the above two statistics, the Benford-specific form of the CDF is unknown.

Test statistics based on the Cramér–von Mises statistic have been introduced in the literature to test discrete distributions, as the Benford one. Following Lockhart et al. (2007), we consider the following forms of the Cramér–von Mises statistic,

$$W_N^2 = N \sum_{d=1}^{8} Z^2(d) t(d), \tag{8}$$

$$U_N^2 = N \sum_{d=1}^{9} (Z(d) - \overline{Z})^2 t(d), \tag{9}$$

$$A_N^2 = N \sum_{d=1}^{8} \frac{Z^2(d) t(d)}{T(d)[1 - T(d)]}, \tag{10}$$

where $t(d) = [P_B(d) + P_B(d+1)]/2$ with $P_B(10) \doteq P_B(1)$ and $\overline{Z} = \sum_{d=1}^{8} Z(d) t(d)$. The main problem of the above statistics, whose asymptotic test values are presented in Table 2, is that only their asymptotic distributions are known. Therefore, their application to small data samples, which is our case, cannot be completely trusted. [A minor point is that their asymptotic distributions are not known in closed form, although a precise method to find them is known and is based on numerical integration (see, e.g., Lockhart et al., 2007)].

In Table 2, we report the test values for the above test statistics for the first-digit distribution of the monkeypox daily case counts in the USA. The null hypothesis of conformance to Benford's law, for the overall monkeypox data, cannot be excluded at 90% confidence level. This result corroborates our finding that monkeypox data comply with Benford's law at a high statistically significance level.

Table 2. Test values for different test statistics for the first-digit distribution of the monkeypox daily case counts in the USA. Also indicated is the asymptotic critical values at 90% confidence level for each statistic and the corresponding reference.

Test	test val.	crit. val.	Ref.
χ_8^2	9.877	13.362	—
D_N^*	0.691	1.012	Morrow, 2014
V_N^*	1.091	1.191	Morrow, 2014
m_N^*	0.795	0.851	Morrow, 2014
W_N^2	0.162	0.351	Lesperance et al., 2016
U_N^2	0.127	0.163	Lesperance et al., 2016
A_N^2	0.732	1.743	Lesperance et al., 2016

Other goodness-of-fit tests could be used, in principle, to test monkeypox data against Benford's law, such as the Goodman's rule of thumb (Goodman, 2016, Campanelli, 2022d), simultaneous confidence intervals for multinomial probabilities [see Lesperance et al. (2016) for a discussion of seven different simultaneous confidence intervals tests], the mean absolute deviation (MAD) criterion (see, e.g. Nigrini, 2012), and the Friedmann test (Giles, 2013), just to cite a few.

However, it is important to stress the fact that, at least to our knowledge, the only Benford-specific test statistic with known analytical expression for the CDF, valid for either small and large data samples, is the Euclidean distance. This makes the Euclidean distance test the ideal one for studying Benford's law in dynamic data, where the number of data points is variable and usually starts from small values.

5. Conclusions

There is an increasing evidence that the number of counts of both death and confirmed cases due to infectious diseases conforms to Benford's law. This is the case, for example, of COVID 19, where extensive analyses have been performed in the last few years. The aim of this paper was to analyze the first-digit distribution of the daily case counts for the ongoing 2022 monkeypox outbreak in the USA.

A global analysis of the data was performed by using 8 different statistical tests, including the "Euclidean distance test", which has been proposed by the author elsewhere to specifically quantify the goodness of fit of a data sample to Benford's law. Our results show that the data comply with Benford's law at a high significance level. This suggests that no manipulations or errors in data collection occurred.

Daily counts of monkeypox cases, and in general death and confirmed cases counts for any infectious disease, evolve in time. In order to follow the spread of monkeypox dynamically, we analyzed the temporal deviation of monkeypox counts from Benford's law. Indeed, a dynamic data analysis of the chronology of the counts could not only flag anomalies but also could frame an anomaly temporally. In the case of monkeypox in the USA, no anomalies were detected, with the temporal series of daily cases conforming to the Benford's distribution to a remarkably high significance level of about 99.96%. The statistical test we used for the dynamic analysis was the Euclidean distance test. The motivation was that, as far as we know, this is the only Benford-specific test with known analytical expression for the cumulative distribution function of the underlying estimator which is valid for either small and large data samples. This last property is strongly required when testing Benford's law in dynamic data, since the number of data points usually starts from small values and than grows in time.

A similar analysis to the one presented in this paper could be applied to both monkeypox counts from other countries when sufficient data become available and/or to future infectious diseases to flag anomalies and fraudulent manipulations either globally or temporally.

References

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

Berger A., Hill, T., Rogers, E., (2009). Benford online bibliography. https://www.benfordonline.net/ (accessed on 2023-09-29).

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

Campanelli, L., (2022c). On the Euclidean Distance Statistic of Benford's Law. *Communications in Statistics - Theory and Methods.* DOI: 10.1080/03610926.2022.2082480.

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

CDC, (2022). U.S. Monkeypox Case Trends Reported to CDC.
https://www.cdc.gov/poxvirus/monkeypox/response/2022/ (accessed on 2022-09-24).

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

Farhadi, N., (2021). Can we rely on COVID-19 data? An assessment of data from over 200 countries worldwide. *Sci. Prog.* 104: 1-19.

Giles, D. E., (2013). Exact asymptotic goodness-of-fit testing for discrete circular data, with applications. *Chilean Journal of Statistics*, 4(1): 19-34.

Goodman, W., (2016). The promises and pitfalls of Benford's law. *Significance* 13: 38-41.

Hill, T. P., (1995a). The significant-digit phenomenon. *Am. Math. Mon.* 102: 322-327.

Hill, T. P., (1995b). Base-invariance implies Benford's law. *Proc. Am. Math. Soc.* 123: 887-895.

Hill, T. P., (1995c). A statistical derivation of the significant-digit law. *Stat. Sci.* 10: 354-363.

Lockhart, R. A., Spinelli, J. J., Stephens, M. A., (2007). Cramér–von Mises statistics for discrete distributions with unknown parameters. *The Canadian Journal of Statistics* 35; 1: 125-133.

Kolmogorov, A., (1933). Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari* 4: 83-91.

Kuiper, N. H., (1960). Tests concerning random points on a circle. *Proc. Koninkl. Nederl. Akad. Van Wettenschappen* A63: 38-47.

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

Leemis, L. M., Schmeiser, B. W., Evans, D. L., (2000). Survival Distributions Satisfying Benford's Law.
Am. Stat. 54: 236-241.

Lesperance, M., Reed, J. W., Stephens, M. A., Tsao, C., Wilton, B., (2016). Assessing Conformance with Benford's Law: Goodness-Of-Fit Tests and Simultaneous Confidence Intervals. *PLos ONE* 11(3): e0151235. DOI:10.137/journal.pone.0151235.

Morrow, J., (2014). Benford's Law, Families of Distributions and a Test Basis. *Centre for Economic Performance.* London.

Nigrini, M. J., (1996). A taxpayer compliance application of Benford's law. *Journal of the American Taxation Association* 18: 72-91.

Nigrini, M. J., (2012). *Benford's Law. Applications for Forensic Accounting, Auditing, and Fraud Detection*, Hoboken, New Jersey: John Wiley & Sons.

Pérez-González, F., Abdallah, C. T., Heileman, G. L., (2007). Benford's Law in Image Processing. IEEE International Conference on Image Processing, 405–408.

Roukema, B. F., (2013). A first-digit anomaly in the 2009 Iranian presidential election. *J. Appl. Stat.* 41: 1, 164-199.

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

Sambridge, M., Tkalčić, H., Jackson, A., (2010). Benford's law in the natural sciences. *Geophys. Res. Lett.* 37: L22301.

Stephens, M. A., (1970). Use of the Kolmogorov-Smirnov, Cramér–von Mises and Related Statistics Without Extensive Tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 32: 115-122.

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

WHO, (2022). https://www.who.int/emergencies/situations/monkeypox-outbreak-2022 (accessed on 2022-09-24).