Journal of Mathematical Finance, 2024, 14, 417-429
https://www.scirp.org/journal/jmf
ISSN Online: 2162-2442
ISSN Print: 2162-2434

Assessing Conformity to Benford's Law with Application to Check China Financial Market

Guojun Fang

Department of Public Teaching, Fuzhou Medical College of Nanchang University, Fuzhou, China
Email: fanggj@163.sufe.edu.cn

How to cite this paper: Fang, G.J. (2024) Assessing Conformity to Benford's Law with Application to Check China Financial Market. *Journal of Mathematical Finance*, **14**, 417-429.
https://doi.org/10.4236/jmf.2024.144024

Received: October 24, 2024
Accepted: November 24, 2024
Published: November 27, 2024

Abstract

Benford's law states that the basic distribution of first digits in naturally produced data is that the frequency of first digits decreases as the digits become larger. This law has been widely used in many areas, such as detecting fraud or manipulation in large datasets. Goodness-of-fit tests are used to assess whether the data obeyed Benford's law. However, conventional statistical tests reject the null hypothesis that data obey Benford's law if the data size is very large. In this paper, we calculated the empirical distribution of first digits of stock close price and daily return in China Stock Market and assessed their conformity to Benford's law by several statistical tests. As a comparison, we introduce the distribution of first two digits as another kind of Benford's law and test whether China Stock Market is close to the law.

Keywords

Benford's Law, Conformity, Goodness of Fit, Null Hypothesis

1. Introduction

Benford's law states that the frequency of first digits tends to be smaller with an increase in the first digits in a naturally produced dataset. This law was first discovered by American astronomer and mathematician Newcomb [1] in 1881. Unfortunately, this discovery did not arouse attention until GE physicist Benford [2] independently rediscovered it in 1938. Thus, this law was named after Benford's. Since then, the law has attracted many different researchers to study it in theory and in practical applications.

Theoretically, this law is so widespread that many mathematicians or statisticians have explained it from basic mathematical theory. Hill's papers [3] [4] provided a rigorous explanation of Benford's law. He presented a strict proof based on several assumptions, including scale invariance, base invariance, and random distributions. However, these mathematical explanations cannot interpret why the law is ubiquitous in many different fields. So, Hill raised an interesting open question: which common probability distributions are following Benford's law? [3].

Leemis *et al.* [5], Engel *et al.* [6], Miller *et al.* [7] [8] studied that the Weibull distribution and the Inverse Gamma distribution are both close to Benford's law. Among other common probability distributions, the Log-normal distribution exists widely in natural phenomena. Fasli and Scott [9], Rodriguez [10], Fang and Qihong [11] showed the distribution is often conforming to Benford's law by using various methods. Since the open question is very important, in his paper Fang [12] named the question after Hill, that is Hill's question and investigated the question for several probability distributions based on a lemma [13].

Based on these researches, Benford's law is often used to check data quality and to discover underlying data manipulation or fraud in large datasets [14]-[16]. The principal idea is, given that some basic conditions are met, genuine data should conform with Benford's law and non-conformity with the law may suggest some data problems have taken place. However, these serious investigations must be grounded on firm statistical testing procedure. So similar to Hill's question, there is still another open question: what are the tests and conditions required to define a dataset following Benford's law?

Data conformity with Benford's law can be tested by using various criteria and statistical tests. So far, several tests are used to assess whether a given dataset conforms to Benford's law, for example, the Z-test, χ^2 test, KS test and MAD test. Among the available tests, Pearson's χ^2 was commonly adopted. However, the adoption of χ^2 and other goodness-of-fit statistical tests tends to reject the null of conformity of Benford's law in a large enough sample size even if the deviations are tiny and unimportant. So, these tests have been criticized in many literatures on Benford's law. Cerqueti and lupi [17] called the problem caused by the large data size as the large n problem. Joenssen [18] performed a comparative study between seven first digit tests for conformity to Benford's law by Monte Carlo simulation. He [19] also tested first two digits for Benford's law besides testing first digits.

Financial market stock exchanging data are very large, which is fit to be used to check large n problem. Meanwhile if we know financial market obeys Benford's law, we can use the law to find out some anomalous behaviors in the markets and detect the reason behind it. But we should know which indices are best indicators to show the market following Benford's law. So far Corazza *et al.* [20] have checked daily distribution of first digits of S&P 500's stock Price and Return against the Benford's law. On the other hand, the market is a desirable experimental subject to find out which is the best test among various existing tests.

In this study, we investigate the large n problem by using China Financial Market stock exchanging data and want to know whether the large n really affects some statistical tests. Meanwhile, we check whether China stock market has the similar Benfordness phenomenon occurring in others financial market, such as S&P 500's stocks. From this check, we filter out the appropriate tests and argue these tests can be used in facing with the large n problem.

The rest of the paper is organized as follows. In Section 2, we introduce the basic materials and methods: Benford's law, Data Collection and Chose tests. Results are presented in Section 3. In Section 4, we conclude the paper with some discussions.

2. Methods

2.1. Benford Law

In this work, whether China Stock Market obeys Benford's law is checked. To achieve this, some selected tests are carefully calculated and compared with other tests. We want to know which statistical test is the best one for assessing conformity to the law in "Big Data". We use the existing R-Benford Tests libraries to make these calculations. Besides, some specific codes for the application are also written in the R language.

How to test a dataset conforming to Benford's law is still an open question. So firstly, we give the definition of the law. But the law has various forms of definition. However, which definition should be chosen to test the benfordness of a dataset is flexible. In this study, we choose the two basic definitions, *i.e.*, the distribution of first significant digit and the distribution of first two significant digit. We state these definition as follows:

$$\text{Prob}\left(D_1 = d_1, D_2 = d_2, \cdots, D_k = d_k\right) = \log\left(1 + \left(\sum_{j=1}^{k} 10^{k-j} \cdot d_j\right)^{-1}\right) \tag{2.1}$$

where $D_1, D_2, \cdots, D_k$ represent the first digits, the second digits…, the kth digits, respectively.

This is a general definition of Benford's law.

Equation (2.1), when applied to the first digit, becomes the basic formula of Benford's law as follows:

$$\text{Prob}\left(D_1 = d_1\right) = \log\left(1 + \frac{1}{d_1}\right), \text{ with } d_1 \in \{1, 2, \cdots, 9\} \tag{2.2}$$

Equation (2.1), When applied to the first two digits is as follows:

$$\begin{aligned}&\text{Prob}\left(D_1 = d_1, D_2 = d_2\right)\\&= \log\left(1 + \frac{1}{10d_1 + d_2}\right), \text{ with } d_1 \in \{1, 2, \cdots, 9\} \text{ and with } d_2 \in \{0, 1, \cdots, 9\}\end{aligned} \tag{2.3}$$

where D_1, D_2 denotes the value of first two digits and $\text{Prob}\left(D_1 = d_1, D_2 = d_2\right)$ represent the probability that the first two digits are equal to $d_1 d_2$.

Equation (2.1), When applied to the Second digits is as follows:

$$\text{Prob}\left(D_2 = d_2\right) = \sum_{d_1=1}^{9} \log\left(1 + \frac{1}{10d_1 + d_2}\right), \text{ with } d_2 \in \{0, 1, \cdots, 9\} \tag{2.4}$$

Table 1 shows the expected frequency of first digits, second digits and first two digits of Benford's law. They are calculated based on Equations (2.2), (2.3), (2.4), respectively.

Table 1. The expected frequency of first digits, second digits and first two digits of Benford's law.

First two digits	0	1	2	3	4	5	6	7	8	9	First digits
1	0.0414	0.0378	0.0348	0.0322	0.0300	0.0280	0.0263	0.0248	0.0235	0.0223	0.3010
2	0.0212	0.0202	0.0193	0.0185	0.0177	0.0170	0.0164	0.0158	0.0152	0.0147	0.1761
3	0.0142	0.0138	0.0134	0.0130	0.0126	0.0122	0.0119	0.0116	0.0113	0.0110	0.1249
4	0.0107	0.0105	0.0102	0.0100	0.0098	0.0095	0.0093	0.0091	0.0090	0.0088	0.0969
5	0.0086	0.0084	0.0083	0.0081	0.0080	0.0078	0.0077	0.0076	0.0074	0.0073	0.0792
6	0.0072	0.0071	0.0069	0.0068	0.0067	0.0066	0.0065	0.0064	0.0063	0.0062	0.0669
7	0.0062	0.0061	0.0060	0.0059	0.0058	0.0058	0.0057	0.0056	0.0055	0.0055	0.0580
8	0.0054	0.0053	0.0053	0.0052	0.0051	0.0051	0.0050	0.0050	0.0049	0.0049	0.0512
9	0.0048	0.0047	0.0047	0.0046	0.0046	0.0045	0.0045	0.0045	0.0044	0.0044	0.0458
Second digits	0.1197	0.1139	0.1088	0.1043	0.1003	0.0967	0.0934	0.0904	0.0876	0.0850	1

Source: Author's calculation with Excel.

We assess conformity to Benford's law based on Equations (2.2), (2.3); *i.e.*, the expected frequency of first digits and first two digits.

2.2. Data Collection

In order to assess conformity to Benford's law, we have collected Stock Exchanging Data published by Chinese companies from the year 2017 to 2021 listed on Shanghai A-Share Market, which were obtained from CSMAR (China Stock Market and Accounting Research Database). From these Trading Data, we choose two indices: Close Price and Daily Return as variables to be used to assess Benfordness of China Stock Market.

Table 2 shows the general information of those Data, including number of Trading Days, number of Companies, number of A-shares and Data size. From the last line, we can see the Data size is very big. Our calculation and analysis are based on these trading Data.

Table 2. The companies listed on Shanghai A-share Market.

Year	2017	2018	2019	2020	2021
Trading Days	224	243	244	243	243
No.of Companies	1396	1450	1572	1800	2037
No.of A-shares	1389	1443	1495	1580	1655
Data Size	298,357	332,246	357,322	369,271	541,618

Source: Shanghai Stock Exchange Statistics Annual (2019) (2022), http://www.sse.com.cn/aboutus/publication/yearly/.

Before our check, the hypotheses are given first. We assumed China Stock Market Obeys Benford's law, that is a data should follow benford distribution. We choose two indices to analyze: Close Price and Daily Return to assess the benfordness. We calculate and assess 2017 to 2021 five years trading Data, respectively. Besides, we also combine these data as a whole to assess benfordness. Specifically, these hypotheses are as follows:

Hypothesis 1: The yearly distribution of first digits of Close Price follows Benford's law;

Hypothesis 2: The yearly distribution of first digits of Daily Return follows Benford's law;

Hypothesis 3: The overall distribution of first digits of Daily Return follows Benford's law.

2.3. Chosen Tests

There are many statistical and non-statistical tests to be used to assess a dataset conforming to Benford's law. We choose the tests based on Joenssen used in his paper [18] and make a comparison between them. According to these tests and comparison, some of tests which may be suitable to assess conformity to Benford's law for large data size are selected.

The common goodness-of-fit test for discrete distributions is the Pearson Chi-Square test which is based on the comparison between observed frequencies and expected frequencies. The following is the calculating formula:

$$\chi^2 = n\sum_{i=1}^{9}\frac{\left(p_i - \pi_i\right)^2}{\pi_i} \tag{2.5}$$

From the above formula, χ^2 is obviously sensitive to the data size n and it may be very large when the data size is big. χ^2 statistic asymptotically follows χ^2 distribution with 8 degrees of freedom. If the value of χ^2 is large, the hypothesis, the data obeying Benford's law should be rejected.

The Kolmogorov-Smirnov test is also commonly used goodness-of-fit test, which is based on the maximum vertical distance between the empirical and expected cumulative distribution functions. By using the supremum function, the test can be computed as follows:

$$D = \sqrt{n}\sup_{j=1,\cdots,9}\left|\sum_{i=1}^{j}\left(p_i - \pi_i\right)\right| \tag{2.6}$$

Similar to the χ^2, if the value of D is too large, the hypothesis is considered to be rejected.

The following test is used by Leemis *et al.* [5] to determine conformance to Benford's law for some parametric survival distributions. The test is based on the distribution of Chebyshev distances between empirical distribution and theoretical distribution under the null-hypothesis. In order to derive asymptotic test statistics, Morrow modified the test statistics used by Leemis *et al.* and by Cho and Gaines as follows

$$m = \sqrt{n}\max_{i=1,\cdots,9}\left|p_i - \pi_i\right| \tag{2.7}$$

Large values of m would reject the hypothesis like the statistical test D. The next test, based on the Euclidean distance between observed and expected frequencies, is also first modified and used by Morrow.

$$d = \sqrt{n\sqrt{\sum_{i=1}^{9}\left(p_i - \pi_i\right)^2}} \tag{2.8}$$

Large distances mean the data seriously deviating from Benford's law. Another test is Freedman's modification of Watson's U^2 statistic for discrete distribution. If Deviation from Benford's law is very large leading to larger U^2, which will reject the null hypothesis.

$$U^2 = \frac{n}{9\cdot 10^{k-1}}\cdot\left[\sum_{i=10^{k-1}}^{10^k-2}\left(\sum_{j=1}^{i}\left(p_i - \pi_i\right)\right)^2 - \frac{1}{9\cdot 10^{k-1}}\left(\sum_{i=10^{k-1}}^{10^k-2}\left(\sum_{j=1}^{i}\left(p_i - \pi_i\right)\right)^2\right)\right] \tag{2.9}$$

J_P^2 statistic is based on the Shapiro-Francia test for normality. The value of J_P^2 is within zero and one. The value closer to 1 means the data closer to Benford's law.

$$J_P^2 = \operatorname{sgn}\left(\sum_{i=1}^{9}\left(\pi_i - \frac{1}{9}\right)\cdot p_i\right)\cdot\frac{\left(\sum_{i=1}^{9}\left(\pi_i - \frac{1}{9}\right)\cdot p_i\right)^2}{\sum_{i=1}^{9}\left(\pi_i - \frac{1}{9}\right)^2\cdot\left(\sum_{i=1}^{9}\left(p_i\right)^2 - \frac{1}{9}\right)} \tag{3.0}$$

Besides the above test, the J-divergence is added to measure the goodness-of-fit between the empirical distribution of first digit and Benford's law. J-divergence is defined as half the sum of two possible Kullback-Leibler distances as follows:

$$KL(P,B) = \sum_{k=1}^{9}\text{Prob}(\text{first digit is }k)\times\log_2\left[\frac{\text{Prob}(\text{first digit is }k)}{B(k)}\right],$$

$$KL(B,P) = \sum_{k=1}^{9}B(k)\times\log_2\left[\frac{B(k)}{\text{Prob}(\text{first digit is }k)}\right]$$

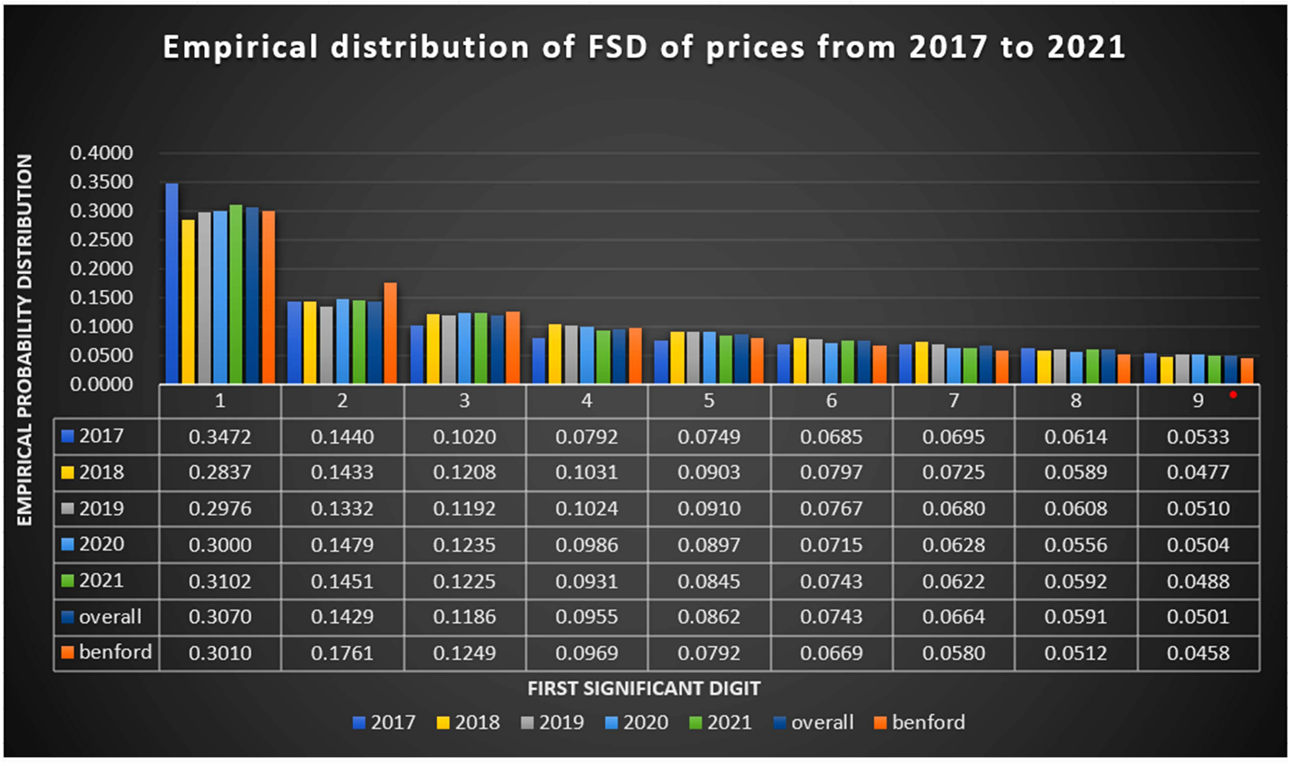
$$J = \frac{1}{2}\times\left[KL(P,B) + KL(B,P)\right] \tag{3.1}$$

B(*k*) denotes the Benford's law expected frequencies. Kullback-Leibler distance, also known as relative entropy, has several good measure-theoretic properties, including non-negative and equaling to zero if and only if P and B are same distributions. The smaller value of J-div means closer to Benford's law.

3. Results

To test Hypothesis 1, the descriptive analysis of China Stock Market Obeying Benford's law is performed by calculating the frequency of first digits of Close Price based on yearly data and overall data. So, we can see six bars plus another bar from Benford's law in **Figure 1**. Obviously, with few exceptions, the height of the bars decreases as the increasing of first digits for each year.

Since the empirical distribution of first digits presents the basic pattern of Benford's law, which intuitively indicates the distribution of first digits of Close Price is almost close to Benford's law. The data columns represent the detailed frequencies of first digits of Closed Price from which the distribution of first digits of each year can be clearly observed. The deviation from the theoretical distribution in 2017 is bigger than the distribution of first digits in other years. The frequency of 1 in 2017 is 0.3472 which is larger than 0.3010. On the contrary, the frequencies of 2, 3, 4 are all very small compared with the expected frequencies.



	1	2	3	4	5	6	7	8	9
2017	0.3472	0.1440	0.1020	0.0792	0.0749	0.0685	0.0695	0.0614	0.0533
2018	0.2837	0.1433	0.1208	0.1031	0.0903	0.0797	0.0725	0.0589	0.0477
2019	0.2976	0.1332	0.1192	0.1024	0.0910	0.0767	0.0680	0.0608	0.0510
2020	0.3000	0.1479	0.1235	0.0986	0.0897	0.0715	0.0628	0.0556	0.0504
2021	0.3102	0.1451	0.1225	0.0931	0.0845	0.0743	0.0622	0.0592	0.0488
overall	0.3070	0.1429	0.1186	0.0955	0.0862	0.0743	0.0664	0.0591	0.0501
benford	0.3010	0.1761	0.1249	0.0969	0.0792	0.0669	0.0580	0.0512	0.0458

Figure 1. Close Price. This is a figure made up of a bar chart and a table. The figure shows the empirical distribution of first digits of Close Price in China Stock Market from 2017 to 2021. From the above bar chart, we can see the empirical distribution of first digit of Close Price is almost close to the distribution pattern of Benford's Law, which can be shown the heights of the bar almost monotonically decrease with the increasing of first digits. Data set columns represent the detailed frequencies of first digits of Close Price and Benford's Law column represents the expected frequency of first digits.

However, the descriptive analysis is not enough, so statistic test procedure is needed. We calculated seven tests listed in the section of Chosen tests and presented their values in Table 3. The values of χ^2, ks, Mdist, Edist, U^2 are all large enough to exceed their critical value. So, we would reject the Hypothesis according to general rule of hypothesis testing. But these large values are caused by the big data size for these tests are all related to the data size n as shown in their calculating formulas. To show the trend clearly, we put the five years data together and repeat the testing procedure. The results are listed in the last row of Table 3. The value of those tests becomes larger than the value of each year. It is obviously caused by larger data size.

Table 3. Tests of conformity of Benford's law for the distribution of first digits of Close Price.

Tests / Year	χ^2	Ks	Mdist	Edist	U^2	J^2	J-div
2017	7820.5	25.192	25.192	35.869	161.95	0.9437	0.0192
2018	5493.8	31.285	18.925	25.745	132.29	0.9788	0.0119
2019	6574.5	31.080	25.639	29.099	131.82	0.9614	0.0140
2020	2780.0	18.645	17.128	19.180	50.01	0.9839	0.0056
2021	4806.7	20.647	22.809	25.905	77.79	0.9773	0.0067
Overall	256474.0	491.770	141.130	212.470	11592.00	0.9743	0.0082

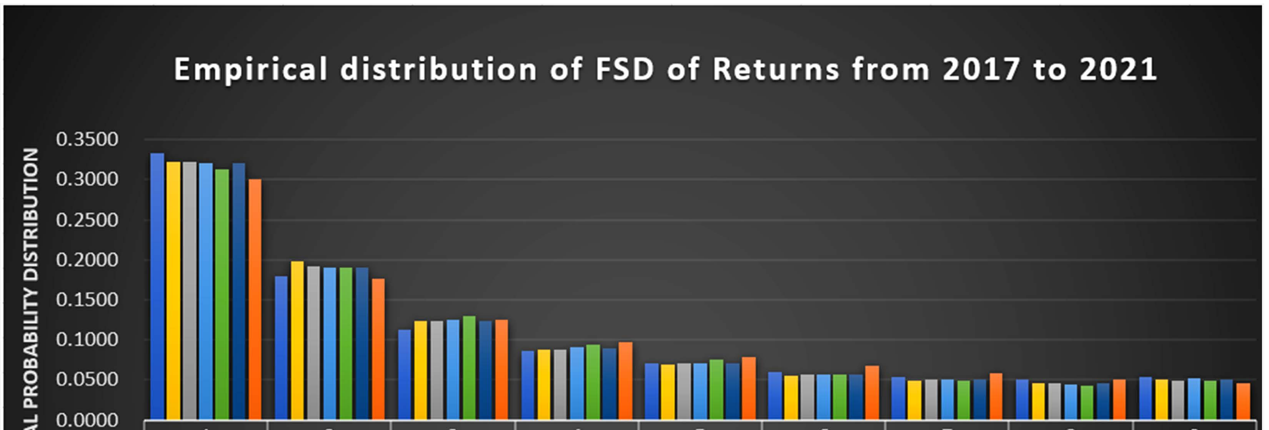
Source: Author's calculation with R.

Even so, these values of tests can also show the deviation between the data and Benford's law for each year. For example, in the year 2017, the value of χ^2 is obviously larger than other years, which shows the data is not much conforming to Benford's law. This corresponds to our previous descriptive analysis. However, the values of J^2 and J-div are not affected by the data size. Thus, these tests better reflect the closeness of Benfordness than other tests.

In the above analysis, we have calculated the Frequency of first digits of Close Price. Although both the descriptive analysis and testing procedure of first digits of Close Price reveal the benfordness of China Stock Market, the closeness is not very good. So, another indicator is introduced. That is Daily Return of Stock. Again, we calculate the first digits of Daily Return.

To test Hypothesis 2, the descriptive analysis of China stock market Obeying Benford's law is performed by calculating the first digits of Daily Return of Stocks based on yearly data and overall data. So we also can see six bars plus another bar from Benford's law in **Figure 2**. Similarly, the height of the bars decreases as the increasing of first digits for each year. Since the empirical distribution of first digits presents the basic pattern of Benford's law, which intuitively indicates the distribution of first digits of Daily Return is almost close to Benford's law and much closer than the distribution of first digits of Close Price. The data columns represent the detailed frequencies of first digits of Daily Return, from which the distribution of first digits of each year can be clearly observed. The deviation from the theoretical distribution in 2017 is bigger than the distribution of first digits in other years, same as the previous analysis. The frequency of 1 in 2017 is 0.3332 which is larger than 0.3010.

Likewise, we calculated seven tests listed in the section of Chosen tests and presented their value in **Table 4**. The values of χ^2, ks, Mdist, Edist, U^2 are all large enough to exceed their critical value, respectively. It is no doubt that the results are related to their calculating formula. However, these values are all small compared to the testing value in analysis of Close Price, which implies Daily Return possibly is a better indicator to show the benfordness of China Stock Market.



	1	2	3	4	5	6	7	8	9
2017	0.3332	0.1788	0.1119	0.0860	0.0712	0.0605	0.0539	0.0510	0.0535
2018	0.3226	0.1974	0.1236	0.0882	0.0685	0.0551	0.0484	0.0457	0.0506
2019	0.3224	0.1913	0.1236	0.0884	0.0708	0.0571	0.0504	0.0466	0.0493
2020	0.3199	0.1909	0.1251	0.0911	0.0707	0.0564	0.0500	0.0442	0.0517
2021	0.3122	0.1909	0.1290	0.0941	0.0747	0.0569	0.0491	0.0433	0.0498
overall	0.3207	0.1902	0.1236	0.0901	0.0715	0.0571	0.0501	0.0457	0.0508
benford	0.3010	0.1761	0.1249	0.0969	0.0792	0.0669	0.0580	0.0512	0.0458

Figure 2. Daily Return. This is a figure made up of a bar chart and a table, which shows the empirical distribution of first digits of daily Return in China Market from 2017 to 2021. From the above bar chart, we also can see the empirical distribution of first digit of Daily Return is almost close to the distribution pattern of Benford's Law, which can be shown the heights of the bar almost monotonically decrease with the increasing of first digits. The below table is the detailed frequency of first digits of Daily Return. Intuitively, compared to Figure 1, this figure shows the distribution of first digits of Daily Return is closer to Benford's Law.

Table 4. Tests of conformity of Benford's law for the distribution of first digits of Daily Return.

Tests / Year	χ^2	Ks	Mdist	Edist	U^2	J^2	J-div
2017	2859.3	21.114	11.632	18.770	144.40	0.9894	0.0065
2018	3877.4	22.198	8.255	18.540	173.72	0.9928	0.0083
2019	3072.8	24.812	7.265	16.061	163.78	0.9957	0.0056
2020	3296.1	24.714	7.678	15.887	153.01	0.9953	0.0061
2021	3931.0	26.666	8.781	16.478	155.50	0.9956	0.0049
Overall	292997.0	532.890	138.480	202.960	16701.00	0.9958	0.0054

Source: Author's calculation with R.

Again, we put these five years data together and repeated the testing procedure. The results are listed in the last row of Table 4. The value of those tests becomes larger than the value of each year. Obviously, it has resulted from the bigger data size.

Even so, these values of tests can also show the deviation between the data and Benford's law for each year, for example, in the year 2017, the value of χ^2 is obviously smaller than other years, which shows the data is much conforming to

Benford's law. This also corresponds to our previous descriptive analysis. However, the values of J^2 and J-div are still not affected by the data size. Thus, again these tests better reflect the closeness of Benfordness than other tests. J^2 is closer to 1 here than the value of the test for Close Price. And, J-div is closer to 0 than the value of the test for Close Price. Those show that Daily Return as an indicator testing for Benfordness is better than Close Price.

Besides, to check the benfordness of China Stock Market further, we tested the conformity to Benford's law by using the distribution of First Two Significant Digits (first two digits) of Close Price as a supplement and a comparison. The result is shown in Table 5. The values of χ^2, Ks, Mdist, Edist, U^2 are all large enough to reject the Hypotheses. The value of χ^2 is larger than the value of χ^2 in Table 3. J^2 is not much close to 1 and is smaller than the value of J^2 in Table 3. So, if testing for conformity to Benford's law using first two digits may be not a good choice compared with using first digits.

Table 5. Tests of conformity of Benford's law for the distribution of first digits of Daily Return.

Tests / Year	χ^2	Ks	Mdist	Edist	U^2	J^2
2017	9617.5	25.980	6.4447	13.613	171.59	0.9342
2018	6978.2	33.565	2.8508	9.417	136.11	0.9614
2019	9350.5	32.366	4.2332	12.187	148.19	0.9291
2020	4255.8	19.299	3.0658	7.973	52.39	0.9708
2021	7205.7	20.647	3.5087	10.879	84.43	0.9635
Overall	261537.0	491.77	15.9630	67.633	16452.00	0.9609

Source: Author's calculation with R.

Similarly, the conformity to Benford's law by using the distribution of First Two Significant Digits (first two digits) of Daily Return is tested as a supplement and a comparison. The result is shown in Table 6. The values of χ^2, ks, Mdist, Edist, U^2 are all large enough to reject the Hypotheses. J^2 is also not much close to 1. Again, this shows first two digits as a kind of Benford's law is not a good choice compared with first digits.

Table 6. Tests of conformity of Benford's law for the distribution of first digits of Daily Return.

Tests / Year		Ks	Mdist	Edist		
2017	6185.0	21.855	4.5095	7.9037	148.11	0.9791
2018	8069.5	24.427	3.7678	8.1888	167.95	0.9781
2019	6964.3	27.527	3.5691	7.6872	160.29	0.9824
2020	10658.0	27.446	5.0129	9.5614	153.62	0.9695
2021	15457.0	31.042	6.0819	11.2290	156.95	0.9684
Overall	313228.0	539.120	21.6670	66.0970	20480.00	0.9778

Source: Author's calculation with R.

4. Conclusions

Assessing the conformity to Benford's law for large data size is still an open question now, especially for the financial market trading data. There are mainly two problems that need to be addressed. One problem is which statistical test is an appropriate choice or a better choice in comparison with other tests. The other is to choose which variable is an indicator to show the benfordness of the big data. Another problem is whether to test first digits or test first two digits or test both. We select China Stock Market trading data as big data here. Six common statistical tests are selected to implement the test procedure; they are χ^2, Ks, Mdist, Edist, U^2, J^2, respectively. Relative entropy J-div is chosen as an addition.

Firstly, descriptive analysis is given before the formal tests based on yearly trading data. We think such a procedure is a necessity to present a visual impression whether the data follow Benford's law. The analysis indeed tells us that China Stock Market is approximately close to the law, whether the indicator is Close Price or Daily Return. This observation is conformed to other financial market such as $&P500. In order to check the observation precisely, we further test the result by using the listed statistical tests. We found these tests (χ^2, ks, Mdist, Edist, U^2) are all sensitive to the data size, except J^2 and J-div. So, we think the J^2 and J-div are fit for such big data test.

Back to the China Stock Market, the Close Price and Daily Return are all to present the benfordness of the Market. But the Daily Return is closer to Benford's law compared with Close Price. Some basic theory behind this is worth to further research, maybe it is related to Stock Price following geometric brown motion. In real application, we think Daily return would be a suitable indicator to test the benfordness of financial market and be used to study other problems, such as system risk.

We know Benford's law has different equations, in this work we have tested the distribution of first digit and first two digits. Obviously, the distribution of first digit is more fit to be used to test the benfordness than the distribution of first two digits. So, our suggestion is, in such kind of tests, testing the distribution of first digit is enough and testing the distribution of first two digits is not necessary, but testing it can be as a supplement.

Acknowledgements

This work was supported by the Science and Technology Research Project of Jiangxi Provincial Department of Education with Grant Number GJJ2203412.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Newcomb, S. (1881) Note on the Frequency of Use of the Different Digits in Natural

Numbers. *American Journal of Mathematics*, **4**, 39-40.
https://doi.org/10.2307/2369148

[2] Benford, F. (1938) The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society*, **78**, 551-572.

[3] Hill, T.P. (1995) A Statistical Derivation of the Significant-Digit Law. *Statistical Science*, **10**, 354-363. https://doi.org/10.1214/ss/1177009869

[4] Hill, T.P. (1995) Base-Invariance Implies Benford's Law. *Proceedings of the American Mathematical Society*, **123**, 887-895.
https://doi.org/10.1090/s0002-9939-1995-1233974-8

[5] Leemis, L.M., Schmeiser, B.W. and Evans, D.L. (2000) Survival Distributions Satisfying Benford's Law. *The American Statistician*, **54**, 236-241.
https://doi.org/10.1080/00031305.2000.10474554

[6] Engel, H. and Leuenberger, C. (2003) Benford's Law for Exponential Random Variables. *Statistics & Probability Letters*, **63**, 361-365.
https://doi.org/10.1016/s0167-7152(03)00101-9

[7] Cuff, V., Lewis, A. and Miller, S.J. (2015) The Weibull Distribution and Benford's Law. *Involve*, *a Journal of Mathematics*, **8**, 859-874.
https://doi.org/10.2140/involve.2015.8.859

[8] Durst, R.F., Chi, H., Lott, A., Miller, S.J., Palsson, E.A., Touw, W. and Vriend, G. (2017) The Inverse Gamma Distribution and Benford's Law.

[9] Scott, P. and Fasli, M. (2001) CSM-349—Benford's Law: An Empirical Investigation and a Novel Explanation.

[10] Rodriguez, R.J. (2004) Reducing False Alarms in the Detection of Human Influence on Data. *Journal of Accounting*, *Auditing & Finance*, **19**, 141-158.
https://doi.org/10.1177/0148558x0401900202

[11] Fang, G. and Chen, Q. (2020) Several Common Probability Distributions Obey Benford's Law. *Physica A*: *Statistical Mechanics and Its Applications*, **540**, Article ID: 123129. https://doi.org/10.1016/j.physa.2019.123129

[12] Fang, G. (2022) Investigating Hill's Question for Some Probability Distributions. *AIP Advances*, **12**, Article ID: 095004. https://doi.org/10.1063/5.0100429

[13] Gauvrit, N. and Delahaye, J.-P. (2018) Scatter and Regularity Imply Benford's Law ... and More.

[14] Nigrini, M.J. (2012) Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection. John Wiley & Sons. https://doi.org/10.1002/9781119203094.

[15] Kaiser, M. (2019) Benford's Law as an Indicator of Survey Reliability—Can We Trust Our Data? *Journal of Economic Surveys*, **33**, 1602-1618.
https://doi.org/10.1111/joes.12338

[16] Li, F., Han, S., Zhang, H., Ding, J., Zhang, J. and Wu, J. (2019) Application of Benford's Law in Data Analysis. *Journal of Physics*: *Conference Series*, **1168**, Article ID: 032133. https://doi.org/10.1088/1742-6596/1168/3/032133

[17] Cerqueti, R. and Lupi, C. (2023) Severe Testing of Benford's Law. *TEST*, **32**, 677-694.
https://doi.org/10.1007/s11749-023-00848-z

[18] Joenssen, D.W. (2014) Testing for Benford's Law: A Monte Carlo Comparison of Methods. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2545243

[19] Joenssen, D.W. (2013) Two Digit Testing for Benford's Law. *Proceedings* 59*th ISI World Statistics Congress*, Hong Kong, 25-30 August 2013, 3881-3886.

[20] Corazza, M., Ellero, A. and Zorzi, A. (2010) Checking Financial Markets via Benford's

Law: The S&P 500 Case. In: Corazza, M. and Pizzi, C., Eds., *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, Springer, 93-102.
https://doi.org/10.1007/978-88-470-1481-7_10