Information Engineering Express
International Institute of Applied Informatics
2020, Vol. 6, No. 2, 20 – 36

Detection of Unnatural Parts of Statistical Data

Tetsuya Nakatoh [*], Takahiko Suzuki [†],
Tsukasa Kamimasu [‡], Sachio Hirokawa [§ ‖ ¶]

Abstract

Ensuring the authenticity of statistical data is important because such data are used for various decision-making tasks. However, in practical applications, several types of data alterations have been reported. Therefore, it is necessary to validate the accuracy of statistical data. Benford's law is a well-known method for detecting unnatural numerical data. According to Benford's law, the occurrence probability of the first significant digits follows a particular distribution. However, the unnatural parts of data cannot be accurately identified. In this study, we attempted to identify the unnatural parts of statistical data available in tabular format. A subset of the target data was specified using the row and column names that define each cell in the table or the words displayed in the table title. By measuring the divergence of the subsets, we identified the unnatural subsets. In this paper, we present the results of the identification of unnatural subsets using the agricultural data acquired from the China Statistical Yearbook.

Keywords: Benford's law, data reliability, statistical data, unnatural subsets.

1 Introduction

The analysis statistical data is essential for correct decision making. Various statistical data analysis methods are used to accomplish this; to guarantee the validity of these analyses, the statistical data must be accurate. However, many types of data alterations have been reported in the real world, with public data being no exception. Therefore, there is a need for a method to detect fraud in statistical data.

Several types of numerical data on the natural and social phenomena are known to follow a specific distribution. Benford's law is a representative example of such distributions [1]. For example, the occurrence probability of the number i appearing in the first significant digits of statistical numerical data, such as the electricity rate of each house, population of each area, and length of each river, is $\log_{10}(1+\frac{1}{i})$, and it has been confirmed

[*] Faculty of Nutritional Sciences, Nakamura Gakuen University, Fukuoka, Japan
[†] Research Institute for Information Technology, Kyushu University, Fukuoka, Japan
[‡] Department of Electrical Engineering and Computer Science, Kyushu University, Fukuoka, Japan
[§] Advanced Institute of Industrial Technology, Tokyo Metropolitan University, Tokyo, Japan
[‖] JVIS R&D Center, Fukuoka, Japan
[¶] Kyushu University, Fukuoka, Japan

to follow a distribution, as shown in the Figure 1. Thus, if certain statistical data do not follow such a distribution, they can be considered to be unnatural. They may represent the results of artificial modifications. The unnaturalness of such data can be evaluated by measuring their degree of deviation using Benford's law.

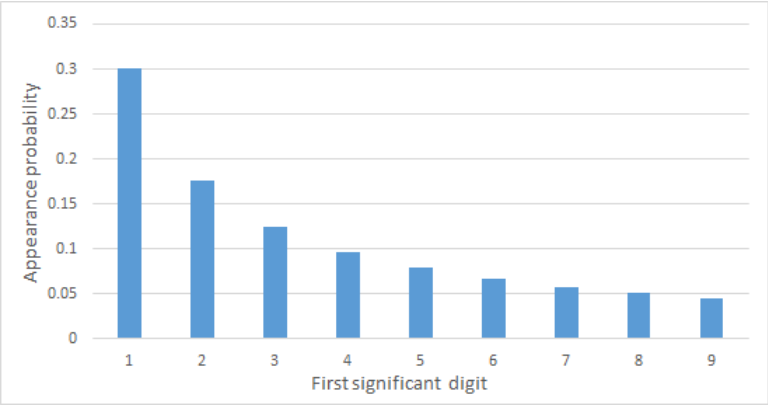
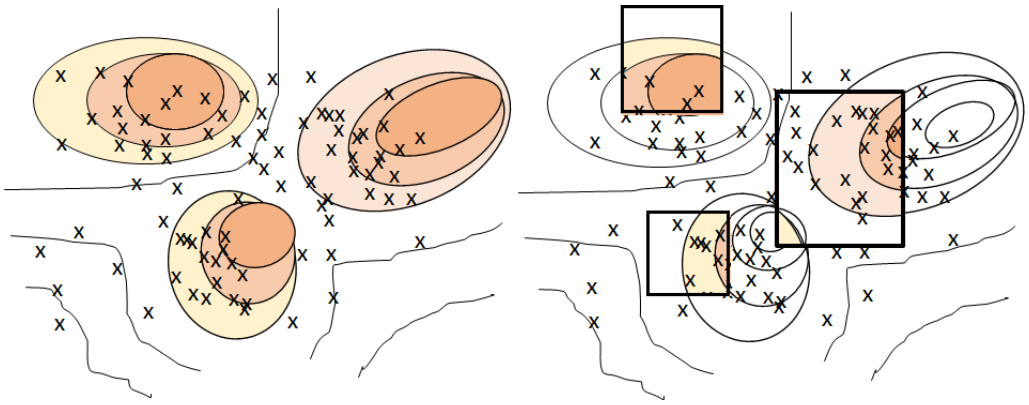


Figure 1: Distribution map of a numerical set according to Benford's law

Unnatural data can be detected using Benford's law; however, their unnaturalness cannot be explained by numerical data. Therefore, it is important to determine a method to interpret such data. Figure 2(a) illustrated a schematic of the unnatural parts of data, denoted by the colored regions. In principle, the unnaturalness of each subset can be evaluated by determining if all subsets of the subject data conform to Benford's law. However, this method has two problems. First, the number of all subsets denotes the order of the exponent of the instances of data, and such evaluations are not practically possible. Second, even if the unnatural parts are determined using this method, they cannot be utilized unless their subsets are interpreted.

In this research, a subset defined by words and attributes is targeted as a subset that can be interpreted by humans. Figure 2(b) demonstrates an example of such subsets, which are enclosed in rectangles. They enumerate the data exhaustively and judge whether they follow a Benford's law distribution. When the numerical value of the target partial data deviates from the distribution of Benford's law, the occurrence of unnaturalness can be represented through words and attributes.

Statistical data are often provided as a series of tables. Each cell can be specified by the title of the table or the items of the rows and columns; correspondingly, the numerical data described in a cell can be obtained. Even when multiple tables are arranged in one table, the value of a particular cell can be retrieved by specifying a partial table. Thus, statistical data can be considered to be multidimensional with table titles and row and column names as arguments. In this study, we constructed a search engine that searches the numerical values appearing in each cell. The index word of each cell is the title of the table including the cell or the item name of a row or column. By designating conditions for specifying a cell for this search engine, it is possible to flexibly specify a set of interpretable numerical values. Accordingly, the unnaturalness can be evaluated by determining the deviation of the distribution of numbers in the first significant digits of each numerical value from Benford's distribution. We applied the proposed method to data related toa agriculture, forestry, and fisheries, obtained from the China Statistical Yearbook, and successfully extracted the unnatural parts.



(a) Ground truth (b) Detection by probing

Figure 2: Probing with condition

2 Method

Section 2.1 summarizes Benford's law and Section 2.2 demonstrates how to extract a subset of statistical data. Furthermore, a proposal analysis system combining them is shown in Section 2.3.

2.1 Benford's Law

Benford's law [12] states that the probability of numerical values in which the first significant digits appear is not uniform; however, it follows a biased distribution in the set of numerical data existing in the natural world. It is generally known that the law is established over a wide range, such as the population of municipalities, traffic volume, length of rivers, and stock prices. The distribution is shown as follows. In the decimal numeric set, the probability $P(n)$ that the first significant digit n is $(1,2,3,4,5,6,7,8,9)$ is given by $P(n) = \log_{10}(n+1) - \log_{10}(n) = \log_{10}(1+\frac{1}{n})$.

The probability that the first significant digit is 1 or 2 is $P(1) = log_{10}(1+1/1) \simeq 0.301$ or $P(2) = log_{10}(1+1/2) \simeq 0.176$, respectively. That is, these two numbers appear in the first significant digit nearly half of all data.

This law was proposed in 1881 by the astronomer Simon Newcomb [11]. It was rediscovered by the physicist Frank Benford [12] in 1938. He verified that the law is valid for various types of data.

2.1.1 Applicability

When the probability distribution of the numerical value of a sample in the logarithmic scale is uniformly spread, as shown in Figure 3, the probability of appearance of the numerical value of the first significant digit corresponds to the area with the width; the ratio of the width is determined by Benford's law formula. Thus, Benford's law holds true for objects that exhibit such probability distributions. Numerical values that increase and decrease exponentially are universally distributed; therefore, they evidently follow Benford's law.

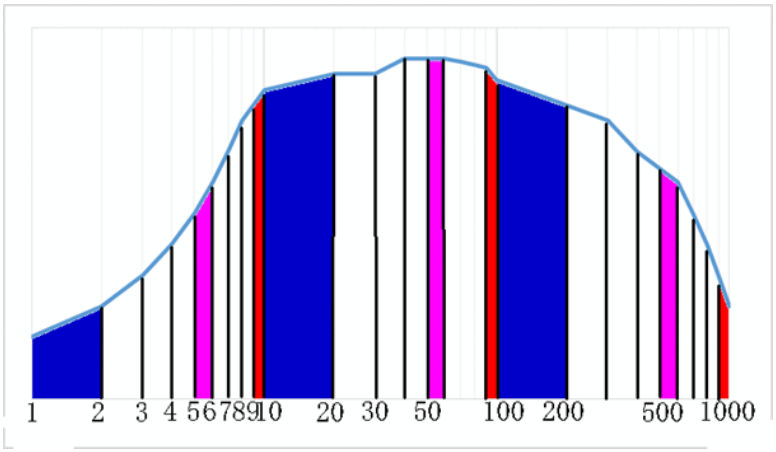


Figure 3: Distribution map of numerical set according to Benford's law

Furthermore, it is also known that various integer sequences, such as the Fibonacci sequence and N! sequence, satisfy Benford's law [19]. Benford [12] showed that this law is valid for 20 numerical sets such as river basins, population of American administrative districts, and physical constants.

However, if the values are distributed within a very narrow range on the horizontal axis of the logarithmic scale, Benford's law cannot be satisfied. This implies that the law cannot be applied to objects whose numerical values lie within a certain range such as the height and weight of adults. Additionally, the law is not valid for intentionally biased data. For example, the annual income of middle-income earners ranges between 3 and 10 million yen; in such cases, Benford's law is evidently not satisfied. Therefore, when Benford's law is used for detecting unnatural data, it is necessary to focus on the nature of the data being used and exclude cases that exhibit a bias in the essential distribution.

2.1.2 Measurement of Divergence Degree

The extent to which the sample distribution deviates from the natural distribution indicated by Benford's law (hereinafter, natural distribution) can be measured by performing the χ^2 test. However, for a relatively small number set, a test method that is more suitable than the χ^2 test is proposed. Max-statistic [8] is a measure based on the largest divergence from Benford's law among the first significant digits ranging from 1 to 9. Distance-statistic [9] is a measure that accumulates the divergence of each digit in the first significant digits.

Let N be the number of samples and $\mathrm{P}(n)$ be the frequency of the number whose first significant digit is n in the sample. Then, χ^2, max-statistic (Hereinafter m), and distance-statistic (hereinafter, d) are represented by the following expressions, respectively:

$$\chi^2 : \sum_{i=1}^{9} \frac{(\mathrm{P}(i) - \log_{10}(1+\frac{1}{i}))^2}{\log_{10}(1+\frac{1}{i})},$$
$$m : \max\{|\mathrm{P}(i) - \log_{10}(1+\frac{1}{i})|, i = 1,2,\cdots,9\}, \text{and}$$
$$d : \sqrt{\sum_{i=1}^{9} (\mathrm{P}(i) - \log_{10}(1+\frac{1}{i}))^2}.$$

The values of m and d on the significance level α [10] are listed in Table 1. The value of α was set to be 0.05 in this study.

Table 1: Values of m and d with respect to significance level α

α	0.1	0.05	0.01
m	0.851	0.967	1.212
d	1.212	1.330	1.569

2.1.3 Application Example

Benford's law is used as a means to verify whether intentional tampering is not performed or false data is not included in the collected data. In 1999, Nigrini [13] focused on Benford's law as a means of finding fraud in accounting audits. Besides accounting audits, it is also used for finding falsification of collected data. For example, Diekmann [16] attempts to detect data tampering by investigating whether statistical indicators of scientific articles follow Benford's law. In addition, Maurus et al. [20] propose an Anomalous-State Detection system that presents the possibility of an attack when the data amount statistic on the network deviates from Benford's law.

2.2 Formalization of Data

Statistical data, such as those contained in the China Statistical Yearbook, are composed of multiple tables. In each column of each table, the elements of each region, occupation, and product are arranged adjacently. For example, the names of places such as Beijing, Heilongjiang, and Henan are arranged in the first column of typical tables. Similarly, elements related to one viewpoint are arranged in each row. For example, if the horizontal axis represents agricultural products, words such as rice, wheat, and corn are in the uppermost row. If each cell in the table represents an area, the chart shows the planted area production amount of each agricultural product in each region. In the chart of "Sown area of farm crops," each cell except for those stated above represents an area. The chart shows the planted area of each agricultural product in each region. In the chart of "Output of major farm products," each cell in the table represents a weight. The chart contains data about the production volume of each agricultural product in each region. We considered such tabular statistical data as records indexed by multidimensional features. The table with viewpoint X on the horizontal axis and viewpoint Y on the vertical axis was considered as a function $T : X \times Y \to R$ from the direct product $X \times Y$ of X and Y to the set of real numbers R.

Given the argument $(x,y) \in X \times Y$, we can find the value of $T(x,y)$ in row x and column y. When statistical data are determined using k types of viewpoints, i.e., $X_1, X_2, \cdots$, and X_k, they can be handled simillarly by considering them as a table of k dimensions instead of a two-dimensional table.

Let c be an element of viewpoint X_p of the statistical data $T : X_1 \times X_2 \times \cdots \times X_k \to R$. The value set $T_X(p)$ of T for c is a set of values defined as $T_X(x) = \{T(x_1, \cdots, x_k) | x_i \in X_k, x_i = c\}$. In other words, it is a set of values obtained by fixing only the p-th argument with respect to c and moving the other parts.

Statistical data(12-8 in Table 3) on the cultivation area of agricultural crops per area are considered to be two-dimensional in the form of $T : Province \times Product \to Area$. At this time, $T_{Province}(\text{Beijing})$ represents the set of areas where various crops are produced in Beijing and $T_{Product}(\text{wheat})$ represents the set of wheat cultivating areas in each region. Given the numerical data of $T_{Province}(\text{Beijing})$, we can verify whether the data on the acreage area in Beijing follows a natural distribution by comparing the distribution of the number of first significant digits with Benford's distribution. Similarly, by analyzing $T_{Product}(\text{wheat})$, we can verify whether the data on the cropping area of wheat follows a natural distribution.

2.3 Exhaustive Evaluation Experiment using Search System

In the conventional application of Benford's law, it is difficult to specify the part of the target data that is inconsistent. In this study, we generated multidimensional tagged numerical values of the numerical value of each cell of the data provided in tabular form. By constructing a retrieval system for these tags, we can extract partial data over several tables accurately and propose an unnatural part detection system based on Benford's law.

An outline of the system is given below.

- The title of the table, attribute name of the row, and attribute name of the column are attached as a tag to each cell (numerical value) in the table. The words in the sentences that explain the table can be used as tags as well.
- Register one cell as one document in the search system. Each tag corresponds to a word in the BoW model.
- Obtain the partial numerical set by specifying the contents to be investigated as a query.
- For the obtained numerical set, test the deviation from Benford's law.

When descriptive text is associated with a table, the words present in the table can be treated as new tags. By repeating the analysis as many times as the number of combinations of tags, a mechanical and comprehensive evaluation can be performed. A similar method of digging down data analysis in units of rows and columns is used in OLAP [7]. We can use any text associated with the data as tags as well as rows and columns.

3 Experiment Data

3.1 China Statistical Yearbook

A statistical yearbook is a comprehensive and systematic description of the basic statistical data gathered from various government ministries. It contains data related to various fields

such as national land, people, economy, and society. It is used as a primary source to understand the state of a country.

In this study, the data from the China Statistical Yearbook [2], compiled and published by the National Bureau of Statistics of China, was targeted. Doubts about the reliability of these data have been expressed occasionally. An article [3] stated that the numerical values in the China Statistical Yearbook are manipulated and released in accordance with the economic growth targeted by the government. In one case [4], inflated numbers were reported by ministries. In addition, according to one report [5], the set of GDP data exhibited an odd distribution that did not comply with Benford's law.

In this study we examined the naturalness of the distributions of agricultural statistical data obtained from the China Statistical Yearbook. The published yearbook is available in Chinese, Japanese, and English in four data formats, i.e., PDF, Excel, JPEG, and HTML (Table 2). The contents of the recorded data are relatively different each year. We used the data on agricultural statistics from Chapter 12 of the 2014 edition, as obtained from [6]. Details of the data used in this study are presented in Table 3.

Table 2: Data format of China Statistical Yearbook available for acquisition

	Chinese	English	Japanese
JPEG	2014–2017	2015–2017	
Excel	2004–2014		1999–2014
PDF	2003		
HTML	1999–2002		

4 Divergence from Benford's Distribution

In this section, we use a few instances of actual data and show analysis examples.

4.1 Agricultural Production in Heilongjiang

We focus on the production volume of the Heilongjiang province, as listed in the 2014 China Statistical Yearbook. Table 4 presents the classification of the production volume of the Heilongjiang province by agricultural crop according to the first significant digit (FSD) of its production volume. The first item in the first row shows that the production of Food/Potatoes is 1.08 million tons and it is distributed under FSD $= 1$.

The frequency of the overall FSD is shown in Figure 4. No significant divergence was observed from the distribution expected according to Benford's law. The corresponding values of m and d are 0.274 and 0.434, respectively, and no significant deviation from Benford's law was found.

4.2 Agricultural Production in Zhejiang Province

Similar aggregate results on the agricultural production in the Zhejiang province are summarized in Table 5 and Figure 5. The corresponding values of m and d are 1.041 and 1.472,

Table 3: Chapter structure of China Statistical Yearbook 2014

Chapter	Title	#
12-1	Agricultural Production Basic Conditions and Sown Area of Farm Crops	100
12-2	Output of Agriculture, Animal Husbandry, and Fishery	336
12-3	Gross Output Value of Agriculture, Forestry, Animal Husbandry and Fishery, and Related Indices	575
12-4	Major Agricultural Machinery at Year-end	348
12-5	Irrigated Area and Consumption of Chemical Fertilizers	348
12-6	Irrigation, Reservoirs, Flood Prevention, Water and Soil Conservation	95
12-7	Water Conservancy Facilities and Area with Flood Prevention Measures by Region (2013)	125
12-8	Sown Areas of Farm Crops	1176
12-9	Planting Structure of Major Farm Crops	212
12-10	Output of Major Farm Products	1407
12-11	Output of Major Farm Products Per Hectare	453
12-12	Output of Major Forest Products	235
12-13	Number of Livestock	492
12-14	Output of Livestock Products	648
12-15	Output of Aquatic Products	752
12-16	Per Capita Output of Major Farm Products	294
12-17	Basic Statistics on State Farms	195
Total	7791	

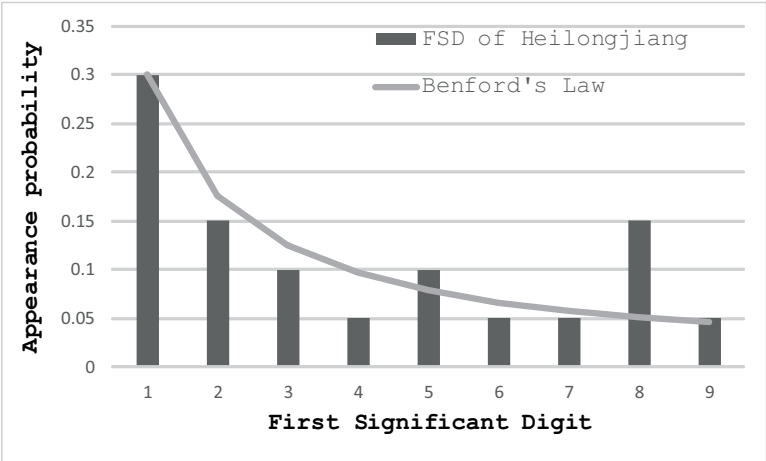


Figure 4: Distribution of FSD on production volume in the Heilongjiang province

Table 4: Distribution of production in the Heilongjiang province

FSD	#	%	Crops (Volume million tons)
1	6	30.0	Sugar beet(1.232), Food/Potatoes(1.08), Oilseed plants(0.19), Fruit/Apple(0.141), Oilseed plants/Oilseed rape(0.001), Oilseed plants/Sesame(0.001),
2	3	15.0	Food/Grain/Rice(22.206), Fruit(2.744), Fruit/Pear(0.028)
3	2	10.0	Food/Grain-Beans/Corn(32.164), Food/Grain/Wheat(0.389)
4	1	5.0	Food/Beans(4.002)
5	2	10.0	Food/Grain(54.959), Silkworm(0.005)
6	1	5.0	Food(60.041)
7	1	5.0	Oilseed plants/Peanuts(0.072)
8	3	15.0	Tobacco(0.089), Fruit/Grape(0.081), Tobacco/Dried tobacco(0.081)
9	1	5.0	Hemp(0.009)

respectively. The distribution of these data deviates significantly from Benford's distribution. By calculating the values of m and d for all provinces, we can determine whether the production distribution follows Benford's law.

4.3 Production Volume of Rice

The process described in Sections 4.1 and 4.2 can be performed on both rows and columns. In other words, the production data for each area of a specific crop can be summarized. Table 6 and Figure 6 present the total amount of rice produced per province.

The corresponding values of m and d are 2.86 and 3.07, respectively. Thus, there is a significant divergence from Benford's law.

5 Result of Analysis

5.1 Overall Result

For the experiment, we used data from six sections of Chapter 12 in the China Statistical Yearbook 2014. When calculating the divergence of the numerical set of each section from Benford's law, no significant results were found in any case (Table 7). In other words, the dataset obtained by integrating the data (data by region, product data) appearing in each section of Chapter 12 is considered to satisfy Benford's law.

Table 5: Distribution of production in the Zhejiang province

FSD	#	%	Crops(Volume million tons)
1	2	9.5	Tea leaves(0.169), Fruit/Citrus(1.93)
2	4	19.0	Food/Grain/Wheat(0.278), Food/Grain-Beans/Corn(0.268), Cotton(0.028), Tobacco(0.002)
3	4	19.0	Fruit/Pear(0.393), Oilseed plants(0.378), Food/Beans(0.338), Oilseed plants/Oilseed rape(0.317)
4	0	0.0	
5	5	23.8	Food/Grain/Rice(5.802), Food/Potatoes(0.523), Silkworm/Silkworm pupa(0.055), Silkworm(0.055), Oilseed plants/Peanuts(0.052)
6	3	14.3	Food/Grain(6.479), Fruit/Grape(0.659), Sweet potato(0.639)
7	2	9.5	Food(7.340), Fruit(7.157)
8	0	0.0	
9	1	4.8	Oilseed plants/Sesame(0.009)

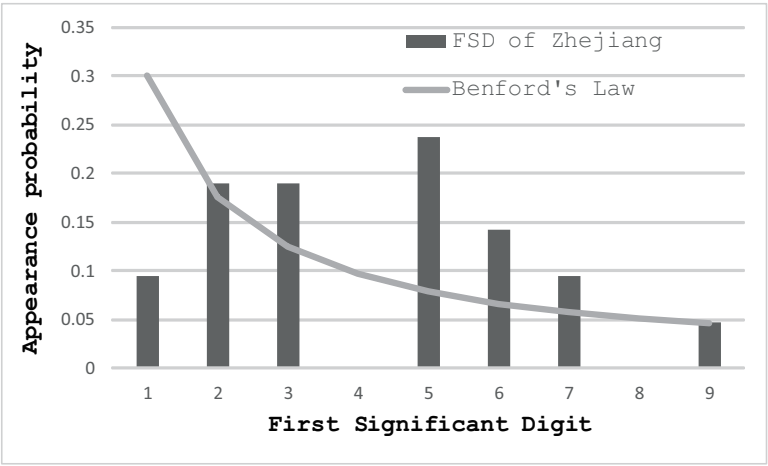


Figure 5: Distribution of FSD on production volume in the Zhejiang province

Table 6: Regional production of rice

FSD	#	%	Province(Volume million tons)
1	10	33.3	Jiangsu(19.2), Hubei(16.8), Sichuan(15.5), Anhui(13.6), Guangxi(11.6), Canton(10.5), Hainan(1.5), Shandong(1.04), Tianjin(0.129), Beijing(0.001)
2	3	10.0	Hunan(25.6), Heilongjiang(22.2), Jiangxi(20.0)
3	2	6.7	Guizhou(3.61), Gansu(0.038)
4	1	3.3	Henan(4.86)
5	8	26.7	Zhejiang(5.80), Jilin(5.63), Liaoning(5.07), Chongqing(5.03), Fujian(5.02), Xinjiang(0.598), Hebei(0.588), Inner Mongolia(0.56)
6	3	10.0	Yunnan(6.68), Ningxia(0.689), Tibet(0.006)
7	1	3.3	Shanxi(0.007)
8	1	3.3	Shanghai(0.868)
9	1	3.3	Shaanxi(0.91)

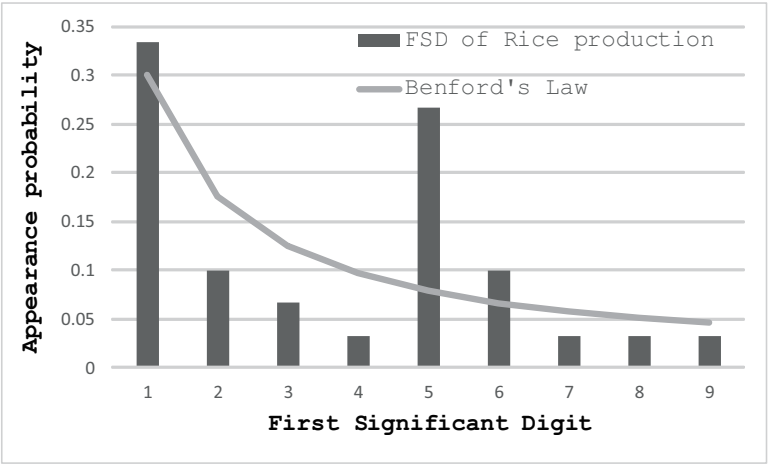


Figure 6: Distribution of FSD on the production volume of rice

Table 7: Test of all numerical data in each section

Item Name	# of Data	m	d
12-8 Sown Areas of Farm Crops	572	0.441	0.739
12-10 Output of Major Farm Products	630	0.440	0.786
12-12 Output of Major Forest Products	97	0.467	0.796
12-13 Number of Livestock	290	0.531	0.921
12-14 Output of Livestock Products	394	0.283	0.529
12-15 Output of Aquatic Products	301	0.682	1.023
Total	2284	0.4532	0.8130

No significant divergence is observed in the values of m and d in this table.

5.2 Crop Cultivation Area

Table 8 presents the calculation results of m and d in terms of acreage of each province for each agricultural product. Individual agricultural products do not diverge from Benford's law. However, the value of d for the total of all agricultural products deviates from Benford's law.

Table 8 lists the calculated values of m and d related to the agricultural product planting areas in each province. Of the 31 provinces, a significant divergence from Benford's law was found in several provinces (m: 7 provinces, d: 4 provinces).

These two results imply the following.

- There is an unnatural part in the data of gross planting area of agricultural products.
- For the planting area of each agricultural product, inconsistency cannot be detected in the distribution of regional data.
- In some areas, there are unnatural parts in the data related to the planting area of agricultural crops.
- The possibility of manipulating the data on crop cultivation area according to crops so that the data on the total planting area of agricultural crops differ from the actual data is staggered in each area.
- Numerical tampering is done independently by each province and there is no evidence that this has been done nationally and uniformly for a certain crop.

The total cultivated area of agricultural products is an important indicator of the local agricultural activities, and it seems that there is sufficient motivation to guarantee that these numerical values are consistent.

6 Related Work

Benford's law was first proposed by Simon Newcomb [11] in 1881. Frank Benford [12] rediscovered it in 1938. Benford verified that this law is valid for various types of data.

Table 8: Cultivation area of each crop

Crop Name	# of Data	m	d
Total cropland planting area	31	0.958	**1.347** $\star$
Food crop planting area	31	0.562	0.870
Food crops / Cereals	31	0.382	0.708
Food crops / Cereals / Rice	30	0.546	0.878
Food crops / cereals / wheat	29	0.236	0.413
Food crops / Cereals - Beans / Corn	31	0.562	0.787
Food crops / Legumes	31	0.396	0.572
Food crops / Potatoes	31	0.575	1.016
Oilseed plants	31	0.254	0.449
Oilseed plants / Peanut	30	0.411	0.718
Oilseed plants / Oilseed rape	27	0.457	0.752
Cotton	25	0.695	0.841
Hemp	21	0.389	0.659
Hemp / Jute	13	0.301	0.547
Sugar	25	0.495	0.735
Sugar / Sweet potato	17	0.942	1.048
Sugar / Sugar beet	9	0.709	0.910
Tobacco	25	0.375	0.753
Tobacco / Dry tobacco	23	0.443	0.779
Vegetables	31	0.696	0.880
Tea plant area	19	0.380	0.587
Orchard area	31	0.538	0.842

Data with $\star$ significantly diverges from the natural distribution at $\alpha = 0.05$

Table 9: Test on Planting Area in Each Province

Province	Number of Data	m	d
Beijing	13	0.382	0.730
Tianjin	13	0.659	1.108
Hebei	20	0.559	0.746
Shanxi	19	0.832	0.913
Inner Mongolia	18	**1.041** $\star$	1.323
Liaoning	18	**1.041** $\star$	**1.547** $\star$
Jilin	16	0.704	1.059
Heilongjiang	18	0.884	1.073
Shanghai	16	0.546	0.772
Jiangsu	18	0.371	0.723
Zhejiang	20	**1.784** $\star$	**1.998** $\star$
Anhui	21	0.857	1.216
Fujian	21	0.573	0.948
Jiangxi	21	0.510	0.728
Shandong	17	0.488	0.858
Henan	21	0.807	1.068
Hubei	21	0.429	0.591
Hunan	21	**1.021** $\star$	1.185
Canton	20	0.554	0.720
Guangxi	21	0.584	0.963
Hainan	18	0.806	1.161
Chongqing	21	0.721	0.981
Sichuan	22	**1.146** $\star$	**1.368** $\star$
Guizhou	20	0.899	1.146
Yunnan	20	0.899	1.250
Tibet	14	**1.212** $\star$	**1.330** $\star$
Shaanxi	20	0.783	0.970
Gansu	20	**1.132** $\star$	1.246
Qinghai	12	0.433	0.849
Ningxia	15	0.576	0.844
Xinjiang	17	0.756	1.111

Data with $\star$ diverge significantly from natural distribution at $\alpha = 0.05$

Nigrini et al. [13] suggested that statistical data fraud can be detected using Benford's law. They conducted several tests on accounting data using Benford's law and demonstrated a method based on Benford's law for detecting numerical tampering.

Holz [14] analyzed the GDP data of China and the USA using Benford's law. The results showed that the real growth rate did not comply with Benford's law for China's GDP data; however, this observation was common for both the USA and China. Thus, Holz argued that it is difficult to find evidence of tampering simply based on the divergence from Benford's law.

Ichinomiya [15] analyzed the domestic researches on statistical data analysis using Benford's law in relation to knowledge gaps in Japan and abroad, along with the differences in previous studies. Ichinomiya also stated that artificially generated fake data can be detected using Benford's law. In addition to the conventional method using the first and second digits, he also confirmed the usefulness of the combination test conducted using the upper and lower two digits from a given set of statistical data.

Andreas [16] performed experiments on a set of fabricated regression coefficients. He found that although the first digits in the fabricated data indicate minimal divergence from Benford's law, the second, third, and fourth digits diverged from Benford's distribution. Arshadil et al. [17] proposed an internet attack detection scheme based on the deviations of the traffic pattern from Benford's law.

IDEA package is used for accounting audits based on Benford's law [18].

7 Conclusion and Future Work

In this study, we built a search engine to analyze numerical data from the China Statistical Yearbook, whose index words include the table titles, row titles, and column names. By variously changing the search conditions used in this search engine, we comprehensively enumerated the interpretable sets of numerical values. This enabled us to evaluate the unnaturalness of the generated subset of statistical data in terms of the divergence of its distribution of the first significant digits from Benford's distribution.

To demonstrate the implementation of the proposed method, we constructed a search engine that examined cells whose values and indices included the area or production volume and the names of regions and products, respectively. A set of cells was considered by fixing its region or product and varying other parameters. We evaluated the naturalness of the subset of numerical data determined by the cell in terms of the divergence of its distribution from Benford's distribution.

Consequently, we found that the data for the seven areas of Inner Mongolia, Liaoning, Zhejiang, Hunan, Sichuan, Tibet, and Gansu significantly deviated from Benford's distribution among the 31 areas in the China Statistical Yearbook.

In this study, we focused on two items to limit the generated subset, namely the area and product. In future, we plan to develop systems that can use more items.

References

[1] Nigrini, M. J., "Benford's Law Applications for Forensic Accounting, Auditing, and Fraud Detection," ISBN: 9781118152850, Wiley, 2012

[2] National Bureau of Statistics of China, "China Statistical Yearbook," http://www.stats.gov.cn/tjsj/ndsj/2014/indexeh.htm, (accessed Jan. 2017)

[3] Nihon Keizai Shimbun, "China's statistics, dubious about reliability," http://www.nikkei.com/article/DGXLASGM19H73_Z11C15A0EA2000/, 2015 (accessed Feb. 2018)

[4] Sankei News: "Liaoning Province, China. Accept false statistics. Fiscal revenue inflated in the past." http://www.sankei.com/world/news/170118/wor170118/0007-n1.html, 2017.1.18 00:40, (accessed Feb. 2018)

[5] Badkar, M., Benford's Law Rises New Doubts About Chinese Economic Data, www.businessinsider.com/benfords-law-questions-chinese-data-2013-1, BUSINESS INSIDER, Jan 11, 2013, (accessed Feb. 2018)

[6] Japan Science & Technology Agency, "Science Portal China," https://www.spc.jst.go.jp/statistics/statictisc_index.html, 2017, (accessed Oct. 2017)

[7] Fraud analysis with SSAS: Benford's law test in OLAP Cubes, www.metrica-bi.de/fraud-analysis-with-ssas-benfords-law-test-in-olap-cubes/, Microsoft, Jun 19, 2015

[8] Leemis, L. M., Schmeiser, B. W, Evans, D. L., Survival Distributions Satisfying Benford's Law, The American Statistician, 54:4, pp. 236–241, 2000

[9] Cho, W., K., T., Gains, B.J, Breaking the (Benford) Law, The American Statistician, 61:3, pp.218–223, 2007.

[10] Morrow, J., Benford's Law, Families of Distributions and a test basis, CEPDP1291, LSE Research Online, http://eprints.lse.ac.uk/60364/ 2010 (accessed Feb. 2018)

[11] Simon Newcomb, "Note on the frequency of use of the different digits in natural numbers," American Journal of Mathematics 4 (1/4), pp.39-–40, doi:10.2307/2369148, 1881.

[12] Benford, F., "The law of anomalous numbers," Proc. of the American Philosophical Society, 78:4, pp.551–572, Mar. 1938.

[13] Nigrini, M., J., I've Got Your Number, Journal of Accountancy, May 1, 1999

[14] Holz, C., A., The quality of China's GDP statistics, In China Economic Review, Volume 30, 2014, pp 309–338, ISSN 1043-951X

[15] Ichinomiya, S., Experimental verification on application of digital analysis, J-STAGE, 2011:21, pp.103–111, 2017

[16] Andreas, D., Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data, Journal of Applied Statistics, 34:3, pp.321–329, 2007

[17] Arshadi1, L., Jahang, A., H., Benford's law behavior of Internet traffic, Journal of Network and Computer Applications, Vol. 40, pp.194–205, 2014

[18] CaseWare Analytics, https://www.casewareanalytics.com/blog/idea-tech-tip-using-benford ’s-law

[19] Sarker, P., B., An Observation on the Significant Digits of Binominal Coefficients and Factorials, Sankhya B.35, pp.363–364, 1973.

[20] Maurus, S., Plant, C., Let’s See Your Digits: Anomalousn-State Detection using Benford’s Law, KDD 2017 Research Paper, Aug. 2017