J.Natn.Sci.Foundation Sri Lanka 2022 50 (2): 387 - 393
DOI: http://dx.doi.org/10.227/jnsfsr.v50i2.10429

RESEARCH ARTICLE

Agricultural statistics

Application of Benford's law in agricultural production statistics

F Hanci
Faculty of Agriculture, Erciyes University, 38280, Melikgazi, Kayseri, Turkey.

Submitted: 24 February 2021; Revised: 17 December 2021; Accepted: 24 December 2021

Abstract: The importance of food supply throughout the world has once again shown its significance in the COVID-19 pandemic period. A continuous food supply is possible with correct agricultural programming. An effective agricultural product programming can only be possible by obtaining precise agricultural data. However, it is very difficult to gather accurate agricultural production statistics from all over the world and confirm their accuracy. In this study, the compatibility of the production statistics of six important agricultural products (wheat, rice, potato, onion, banana, apple) which had been collected from local sources, and had published as open-source by the Food and Agriculture Organization of the United Nations, with Benford's law was examined for the first time. Data for the last two decades are used to ignore the impact of annual fluctuations. The compatibility of theoretically expected and observed data was tested by Chi-square (χ^2) and Mean Absolute Deviation (MAD) tests. Although inconsistencies were found in some data by examining the numbers in the first, second, and first two digits, in general, the MAD test results gave a mostly concordant result.

Keywords: Agriculture, food, nutrition, production, statistics.

INTRODUCTION

In the globalizing world, it has become impossible to consider the agricultural production of any country independently from other countries. In particular, the Covid-19 pandemic has demonstrated the importance of agricultural production and the food supply chain globally. Accordingly, the reliability of statistical data on world agricultural production has started to be on the agenda again. The largest agricultural production statistics data available worldwide is published by FAOSTAT (FAO, 2021). These data form the basis for many scientific studies as well as for future planning. The reliability of these data is primarily the responsibility of each country. A large number of models and methods can be used to prevent the risk of fraud and to detect fraud occurring in published large datasets. One of these methods is Benford's Law (Benford, 1938). Benford's Law, which is a proactive method, determines the numerical frequencies in the digits of the dataset. The direction and magnitude of deviations between the observed value and the rates of Benford's Law can give an idea about the reliability of the dataset. Although the data with deviations are not definitive evidence of fraud, the researcher can create the audit plan by considering the deviations in the dataset. The Benford's Law hypothesis is based on an article by Simon Newcomb published in the American Journal of Mathematics, about the incidence of numbers in digits (Newcomb, 1881). There were no calculators at the time, and many complex mathematical calculations could be made using logarithm tables printed on paper. Newcomb noticed that the first pages of logarithm tables are more obsolete than the last pages. Also, it has been found that the numbers starting with "1" are used more by the researchers than the numbers starting with "2", and those starting with "2" are used more than starting with

* Corresponding author (fatihhanci@erciyes.edu.tr; https://orcid.org/0000-0002-2015-0351)

"3". Newcomb transformed this research into a formula (Lemis *et al*., 2000):

Probability $= log_{10}(1+1/d), (d = 1,2,3,4,5,6,7,8,9)$ (*first digit number*)

According to Newcomb's research, the probability of "1" in the first digit is 0.3010 while the probability of being in the second digit is 0.1139 (Miller, 2015).

Frank Benford, a physicist in General Electric's laboratory in New York, published the results of his study in 1938, in which 20 different groups were examined with a dataset totaling 20,229 (Benford, 1938). 30.6% of this data started with number "1"; 12.4% with "3"; 8% with "5"; and 4.7% with "9" in the Benford observation. Benford's analysis shows a logarithmic distribution rather than a homogeneous distribution (Akkaş, 2015). The logarithm functions of Benford's Law are as follows (Ertikin, 2017):

For the first digit of the numbers;

$$P(D_1 = d_1) = log(1+ (1\ d_1); d_1 \in (1,2,3, ..., 9) \quad ...(1)$$

For the second digit of the numbers;

$$P(D_2 = d_2) = \Sigma\ log(1+ (d_1 d_2); d_2 \in \{0,1,2,3, ...,9\} \quad ...(2)$$

For the first two digits of the numbers;

$$P(D_1 D_2 = d_2 d_2) = log(1+ (1+ 1\ d_1 d_2); d_1 d_2 \in (10,..,99) \quad ...(3)$$

For example; The probability that the number in the first digit of a number is "2" and the number in the second digit is "8" can be calculated by the following formula:

$$P(28) = Log_{10}(1 + 1/28) = Log_{10}(29/28) = 0.01524. \quad ...(4)$$

Benford's Law has been used by researchers from many different disciplines since its inception. While these researchers initially focused on the mathematical explanation of the law, in the following periods, research studies on different fields such as COVID-19, social welfare programs, and academic fraud were emphasized (Horton *et al*., 2020; Lee *et al*., 2020; Azevedo *et al*., 2021). Various prerequisites must be met for a dataset to yield results following Benford's Law (Nigrini, 2000). The dataset should describe the magnitude of similar phenomena. For example, the population of cities, lengths of rivers, stock values, or daily sales amounts. Also, the values in the dataset should not have an upper or lower limit. For example, daily working hours do not follow Benford's Law, as they have to take a limited value between "0" and "24". The values in the dataset should not be specified numbers. Therefore, data consisting of determining numbers such as citizenship identification number, tax number, credit card number, or telephone number are not data by Benford's Law. The Benford's Law analysis technique consists of two stages: general analysis and special analysis tests. A general analysis test gives an idea about the data. These tests are the first digit and second digit tests. Special tests are the first two digits, the first three digits, the last two digits, and duplicate recording tests (Yanık & Samancı, 2013). First and second digit tests cannot be used for sampling in control. However, the second digit test can easily detect fundamental abnormalities in the data (Goh, 2020).

In this study, the compliance with Benford's Law of the data published by FAOSTAT, showing the production quantities for six agricultural products, has been investigated.

MATERIALS AND METHODS

The sample data consists of six agricultural products (wheat, rice, potatoes, onion, banana, apple) published during 2000-2019. In addition to these six species, to provide an overview of all agricultural production statistics and to reach a cumulative result, the values created by bringing together all the data were also examined. Production amounts (tons) in the last two decades were collected on a country basis. In the selection of products, both the strategic importance and the production potentials in different geographies of the world are taken into consideration. For this purpose, a total of 14907 data elements including 2461 wheat, 2308 rice, 3095 potatoes, 2769 onions, 2392 bananas, and 1882 apples were examined. The frequency of the first, second, and first two digits in these data was determined using the Microsoft Excel Office program.

After calculating the frequencies of the numbers in the analysed dataset, it is necessary to determine the "conformity" limits of the deviations according to Benford's Law. At this stage, conformity tests are used to determine how much deviation there is between the observed rates and Benford's Law rates, and whether this deviation is significant. Conformity tests used in Benford's Law include the Z-Statistics Test, Chi-square (χ^2) Test, Kolmogorov-Smirnov Test, and Mean Absolute

Deviation Method (MAD). Although the Z-Statistics Test, χ^2 test, and Kolmogorov-Smirnov Test are based on statistical foundations, they can often be affected by the number of data. In this study, two methods were used to confirm the theoretical expectation of Benford's Law: MAD and χ^2 test. The MAD method often gives results open to interpretation, but they are independent of the number of data. Besides, this test method has a very practical application area (Druica *et al*., 2018). The MAD score is defined as the mean of the absolute value of the difference between the frequency of each first digit within the sample, and the frequency as determined by the formula:

$$MAD = \sum i \frac{|\mathrm{A}f - \mathrm{E}f|}{K} \quad ...(5)$$

where *Af* is the actual frequency of the leading digit observed, E*f* is the expected frequency as determined by Benford, and *K* is the number of leading digit bins (equal to 9 for the first, 10 for the second, and 90 for the first two leading digits).

The second measure of conformity is the χ^2 test statistic which compares expected frequencies and observed frequencies in one or more categories of a contingency table for leading digits 1–9. χ^2 is calculated as follows:

$$\chi^2 = \sum i \frac{(\mathrm{A}v - \mathrm{E}v)^2}{\mathrm{E}v} \quad ...(6)$$

A*v* is the actual value of the leading digit observed, E*v* is the expected value as determined by Benford.

The Microsoft Excel Office program was also used in these analyses. To evaluate the results obtained from the χ^2 test, a table of critical values of distribution with d degrees of freedom was used.

RESULTS AND DISCUSSION

The results of the first digit analysis of selected crops are shown in Table 1. To apply the χ^2 test, null hypotheses (H_0: the differences between the expected and observed values are small enough to be considered insignificant) were set and the significance level approved at 5%. To interpret the results, Chi squared critical values tables were used. Results below the threshold value specified in this table prove "compatibility," and χ^2 test results above this threshold value emphasize the inconsistency. According to the Chi-square critical values table, the critical value found at 8 and 9 degrees of freedom and 5% significance level is 15.507 and 16.919 respectively. The hypothesis in which the number set of six agricultural products conforms to Benford's Law has been rejected according to the χ^2 test. However, it should not be forgotten that the χ^2 test can be directly affected by the number of samples in the dataset, and tends to reject the null hypothesis even for small departures from the expected distribution.

Table 1: Results of the first digit analysis

Number	Wheat		Rice		Potato		Onion		Banana		Apple		Complete	
	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.
1	741	823	772	802	1075	912	816	857	740	686	540	495	3921	4575
2	433	455	427	426	571	532	476	417	360	433	341	380	3010	2643
3	307	230	216	279	374	386	345	401	346	346	272	221	1751	1863
4	238	200	188	192	257	318	285	220	190	160	126	224	1774	1314
5	195	165	155	141	189	265	237	239	206	170	134	153	1212	1133
6	165	151	142	144	193	229	205	187	162	193	152	153	1212	1057
7	143	153	143	96	129	158	141	179	155	155	122	108	855	849
8	126	160	150	132	177	161	144	157	136	149	117	74	586	833
9	113	124	116	96	129	134	120	112	97	100	79	74	586	640
Σ	2461		2308		3095		2769		2392		1882		14907	
χ^2_8	52.796		42.550		87.627		46.250		37.434		114.950		414.403	
MAD	0.014		0.014		0.004		0.011		0.013		0.014		0.004	

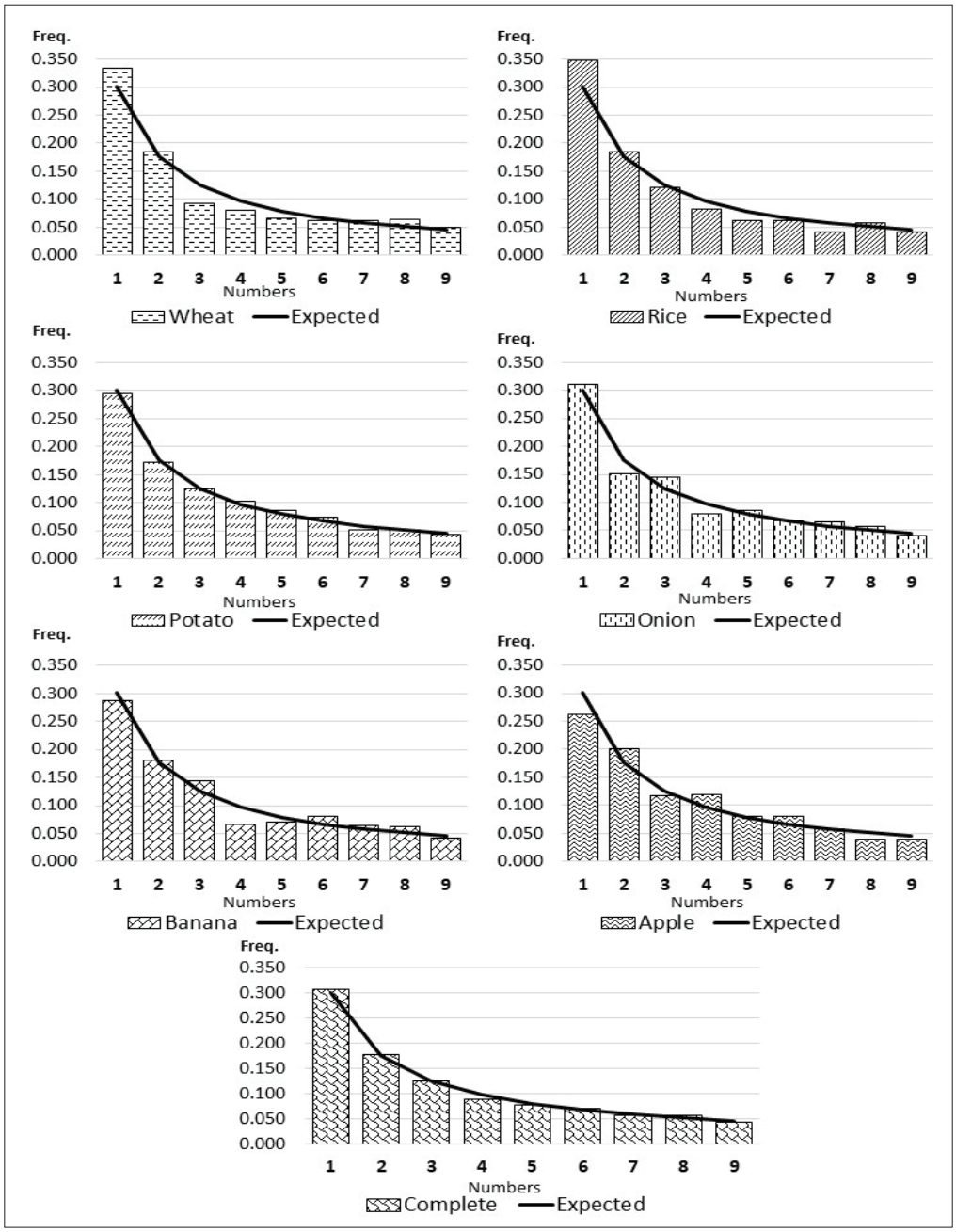

Figure 1: Relative frequencies of the first digits

The MAD test is more robust since it ignores the number of records. The lower the MAD, the smaller the average difference between the observed and theoretical distributions. MAD values under 0.015 suggest conformity (Silva and Filho, 2020). In considering this criterion, it may be concluded that the investigated agricultural data are accordant with the Benford distribution for all crops. When Table 1 and Figure 1 are examined, it is seen that the most accordant distribution in the first digit is observed with potato, according to Benford's law. It was followed by a dataset in which all products were displayed on a single list.

According to the MAD values, the closest values to the expected frequencies were observed in the dataset of the potato. The MAD value in the list where all products are listed together is also similar to potatoes. In onion, the data corresponding to a first digit value of "2", "3", and "4" showed more deviation than expected. A similar situation is found in the data starting with "1", "5", and "7" in rice.

The distribution of the numbers in the second digits is shown in Table 2 and Figure 2. When the compatibility of these distributions with Benford's Law is checked with χ^2, it is seen that the potato, onion, and apple results are compatible at different probability levels ($p = 0.1348$, $p = 0.0236$, $p = 0.0167$ respectively). In the MAD analysis results, the data of all products were found to be consistent.

Table 2: Results of the second digit analysis

Number	Wheat		Rice		Potato		Onion		Banana		Apple		Complete	
	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.
0	294	387	276	343	370	403	331	391	285	386	225	272	1781	2182
1	279	265	262	268	352	339	315	316	272	286	214	231	1695	1705
2	267	251	251	261	336	342	301	312	260	245	205	200	1620	1611
3	256	257	240	225	322	308	289	262	249	201	196	181	1553	1434
4	246	248	231	233	310	291	278	247	239	198	189	179	1493	1396
5	237	237	223	201	299	286	268	279	231	247	182	204	1439	1454
6	229	236	215	182	289	313	258	257	223	207	176	160	1390	1355
7	222	174	208	207	279	266	250	229	215	216	170	159	1345	1251
8	215	200	202	200	271	250	242	232	209	213	165	152	1303	1247
9	209	199	196	184	263	293	235	243	203	186	160	144	1265	1249
Σ	2454		2304		3091		2768		2385		1882		14884	
χ^2_9	43.325		25.907		13.665		19.996		57.043		20.207		115.840	
MAD	0.006		0.007		0.006		0.007		0.011		0.009		0.006	

Table 3: Summarized results of the first two-digit analysis

	Wheat	Rice	Potato	Onion	Banana	Apple	Complete
Σ	2454	2304	3091	2768	2385	1882	14884
MAD	0.002	0.002	0.001	0.002	0.002	0.003	0.001
χ^2_{89}	221.387	222.031	110.174	180.380	216.021	208.868	222.789

When all six products are listed together, it can be said that the data with "0" in the second digit are doubtful in terms of compliance with Benford's law. When the graphs of the expected and observed frequencies in Figure 2 are examined, it is understood that the largest deviation is in the values ending with "0" and "5". But graphs of the relative frequencies of the first digit do not exhibit such behavior. The most obvious difference between these two measurement systems is that it is not possible to display a value starting with the number "0" in the first digit analysis. The discrepancy between "0" and "5" in the second digit analysis may be due to local authorities' preference to round data associated with "1", "4" and "6" to "0" or "5".

In evaluations using Benford's law, interrogation of the first two digits is generally used in complex data. In this study, all six agricultural products were examined separately and as a single list according to their numbers in the first two digits. Since there are a total of 90 (values between 10 and 99) categories, the relevant results are shared in the summary table.

The results obtained according to the numbers in the first two digits are parallel to the results in the second digit. According to MAD results, potatoes and the total list were found to be the most compatible with Benford's law. In the χ^2 test, only the value obtained from potatoes was found to be compatible with Benford's law (p value: 0.9364).

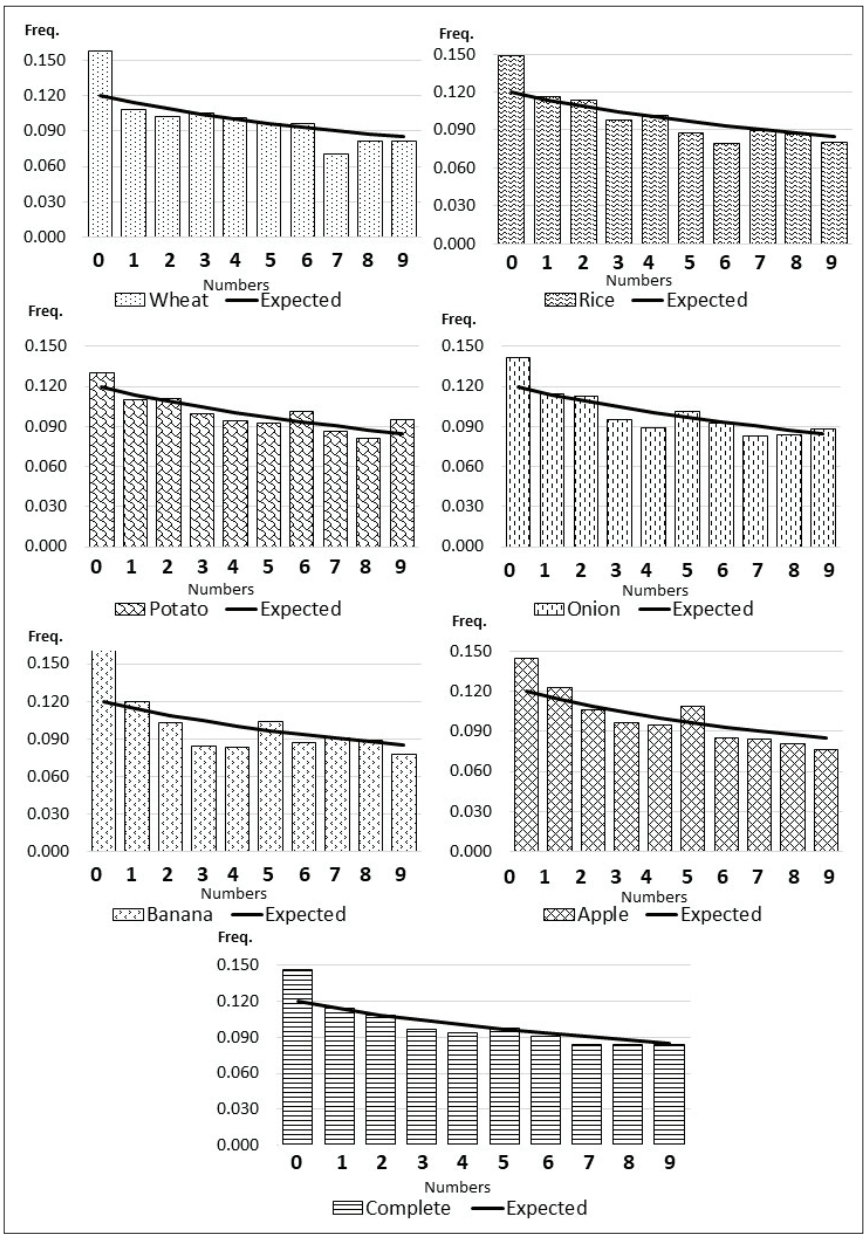

Figure 2: Relative frequencies of the second digits

CONCLUSION

As mentioned above, collecting precise data on agricultural production is a very difficult process. Governments take various measures to increase the reliability of this data. Mostly, local organizations of the central government are trying to create this data set by conducting on-site investigations. However, it is often impossible to achieve exactly this. In some cases, the statements of the farmers are taken as a basis, and correct notifications are tried to be made with various incentives or penalties. The control of the combined data by the central authorities is generally attempted by comparison with the previous year's data, comparison with commercial records, or on-site inspections by assigned inspectors. However, these measures are often insufficient. Benford's Law is a practical statistical instrument for specifying doubtful activities in data. The agricultural production of a particular region depends on very uncertain factors such

as low temperature, drought, and excessive precipitation. Benford's Law is not a system that takes these factors into account directly. Even under uncertain conditions, it is a control system created according to normal distributions in data sets that are formed without human intervention. Therefore, regardless of the conditions, the reliability of the obtained data sets can be verified with Benford's Law. The arguments presented here do not constitute proof of illegality, but some suspected data were identified as consistent departures of agricultural statistics' reports from NBL. Benford's Law has been tested for the first time in agricultural statistics reported by all countries of the world. Therefore, the findings obtained in this study should be evaluated at the "giving an idea" stage. Also, it should not be forgotten that collecting such agricultural data on-site is a very difficult task. However, no major deviations were found in the compatibility analysis with MAD tests. This situation proves the reliability of the data obtained from FAOSTAT despite all difficulties. In this study, "doubtful" numbers in different digits can give local data collectors a quick control chance. In future studies, in addition to the six products discussed in this study, the evaluation of different agricultural products may offer new opportunities to the commentators. Also, agricultural statistics monitoring on a country basis may yield useful results.

Conflict of interest

The author declares that there is no conflict of interest.

REFERENCES

Akkaş M.E. (2015). Testing distribution of gold returns by Newcomb-Benford law. *The Journal of International Social Research* **8**(40): 577–584. DOI: https://doi.org/10.17719/jisr.20154013940

Azevedo C.S., Goncalves R.F., Gava L.V., & Spinola M.M. (2021). A Benford's Law based methodology for fraud detection in social welfare programs: Bolsa Familia analysis. *Physica A: Statistical Mechanics and its Applications* **567**: 1–15. DOI: https://doi.org/10.1016/j.physa.2020.125626

Benford F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society* **78**(4): 551–572.

Druica E., Oancea B., & Valsan C. (2018). Benford's law and the limits of digit analysis. *International Journal of Accounting Information Systems* **31**(1): 75–82. DOI: https://doi.org/10.1016/j.accinf.2018.09.004

Ertikin K. (2017). Fraud auditing: a services business example for computer-aided Use of Benford's law. *The World of Accounting Science* **19**(3): 696–726. DOI: https://doi.org/10.31460/mbdd.609957

FAO (2021). Food and Agriculture Organization of the United Nations, Statistics Division. Available at *http://www.fao.org/faostat/en/#home*, accessed 28 January 2021.

Goh C. (2020). Applying visual analytics to fraud detection using Benford's law. *Journal of Corporate Accounting Finance* **31**: 202–208. DOI: https://doi.org/10.1002/jcaf.22440

Horton J., Kumar D.K., & Wood A. (2020). Detecting academic fraud using Benford law: The case of Professor James Hunton. *Research Policy* **49**(8): 1–19. DOI: https://doi.org/10.1016/j.respol.2020.104084

Lee K., Han S., & Jeong Y. (2020). COVID-19, flattening the curve, and Benford's law. *Physica A: Statistical Mechanics and its Applications* **559**: 1–12 DOI: https://doi.org/10.1016/j.physa.2020.125090

Lemis L.M., Schmeiser B.W., & Evans D.L. (2000) Survival Distribution Satisfying Benford's Law. *American Statistician* **54**(4): 236–24. DOI: https://doi.org/10.2307/2685773

Miller S.J. (2015). A quick introduction to Benford's law. In: *Benford's Law: Theory and Applications.* Princeton University Press, USA.

Newcomb S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics* **4**(1): 39–40.

Nigrini M. (2000). Digital analysis using Benford's law: tests and statistics for auditors. EDPACS **28**(9): 1–2. DOI: https://doi.org/10.1201/1079/43266.28.9.20010301/30389.4

Silva L. & Filho D.F. (2020). Using Benford's law to assess the quality of COVID-19 register data in Brazil. *Journal of Public Health* **43**(1):107–110. DOI: https://doi.org/10.1093/pubmed/fdaa193

Yanık R. & Samancı T.H. (2013). Benford's law and a practical implementation in public sector about its application to accounting data. *Journal of Graduate School of Social Sciences* **17**(1): 335–348.