EKONOMSKE TEME **(2013) 51(3): 515-524**

http://www.eknfak.ni.ac.rs/src/Ekonomske-teme.php

EXAMINING RELIABILITY OF LARGE FINANCIAL DATASETS USING BENFORD'S LAW

Konrad Grabiński
Cracow University of Economics, Poland
✉ *kg@uek.krakow.pl*

Zbigniew Paszek
Andrzej Frycz Modrzewski Kraków University, Poland
✉ *zpaszek@afm.edu.pl*

UDC
336:51-7
Preliminary paper

Abstract: In the article, the authors have analyzed large financial datasets from the perspective of consistency with Benford'a law. Two groups of datasets have been investigated: selected accounting items reported by European publicly listed companies and profitability ratios. We argue that if datasets representing components of financial ratios are reliable from Benford's law perspective, also datasets representing financial ratios build on these accounting items are reliable. Presented study provides evidence that if large financial datasets are of high quality, also datasets constructed on previous datasets are of high quality.

Received:
07.07.2013.
Accepted:
14.10.2013.

Key words: Benford's law, profitability ratios, reliability of financial data

1. Introduction

Benford's law is defined as a law of leading digit and states that in the randomly collected large dataset of numbers, the frequency distribution of digits is not equally likely. To be more specific, the probability that in randomly selected number, the first digit is "1" is about 30%, while the probability that the first number is "9" is only 5% (Orita et al., 2010, p.328). The first digit in a given number is called a leading digit, and the probability of its occurrence is calculated as follows:

$$P_{d_1} = \log_{10}\left(1 + \left[\frac{1}{d_1}\right]\right)$$

where: d_1 =1,2,…,9

P_{d1} – probability of the digit d_1

This simple formula allows calculating probability of occurrence of the first digit in randomly selected number from large dataset of numbers. Origins of the discovery can be traced back to 1881 when mathematician and astronomer Simon Newcomb published short article about the digit distribution in natural numbers (Newcomb, 1881). In 1938, the physician Frank Benford (1938) published the article in which he formulated a law known henceforth as Benford's law. Most of empirical datasets gathered in various fields of study are consistent with Benford's law regardless of the measurement unit used. For example, in financial and accounting datasets, it is not important what currency is used (i.e. euro or US dollars) – Benford's law is working properly (Lobert, 2008, p.104). Some argue that Benford's law is a natural phenomenon and can be considered as a signature of the nature (Bhattacharya et al., 2011, p.577). Ch. Breunig and A. Goerres (2011, p. 534-545) have performed research on democratic elections in Germany for the 1990-2005 period. They concluded that there is no basis for question reliability of democratic elections.

As a result Benford's law can be applied as a measure of reliability of a given dataset. When given dataset is not consistent with Benford's, it can be interpreted that some numbers of this dataset might have been manipulated or distorted in other way. From this perspective, some properties of Benford's law are very useful for detecting accounting frauds and errors. Nowadays, new tools and procedures are being implemented in financial audit, what was proposed by M. J. Nigrini in 1993. However, it is important to notice, that when digit distribution is not consistent with Benford's law it doesn't automatically mean that a fraud or an error have been detected. Some datasets are biased, for example when analysis is performed with regard to sales invoices and usually a company sells small quantities of stock, whose value is between 50 USD and 100 USD (i.e. gas stations).

The aim of the article is to analyze large financial and accounting datasets with regard to (1) values of accounting items derived from balance sheet and profit and loss account and (2) financial ratios' values of European listed companies. The analysis is limited only to profitability ratios: return on asset, return on equity and return on sales. As accounting items are concerned, the analysis is limited to these items, which are used to calculate profitability ratios. The main thesis of the article states that if components of financial ratios are consistent with Benford's law also financial ratios build on these components should be consistent with Benford's law. The result of the analysis may be important for users of financial reports, especially when financial analysis is concerned.

2. Research Design

The sample is derived from Amadeus database and is limited to financial reports of European listed companies. Accounting of listed companies is

perceived to be of the highest quality and is a subject to public supervision through independent financial audit. Additionally, stock exchange institutions impose on publicly listed companies disclosure policies and severe penalties in case of infringement. In most cases publicly listed companies are the biggest and the most important in the economy.

The initial sample consisted of 12 466 European public companies for the time period of 2003-2012, what as a result provided 124 660 firm-year observations. The final sample is smaller due to missing data in accounting items or years, (Table 1). The following accounting items have been analyzed: net profit, equity, sales, total assets and henceforth are called the first category datasets in this analysis. These accounting items are the most important reporting financial figures or at least are among the most important. Additionally, they are widely used in many financial ratios. In particular they are used to calculate profitability ratios: return on assets, return on equity and return on sales, these data also have been downloaded from Amadeus database. Henceforth, data representing profitability ratios are called the second category datasets.

Table 1 Basic Statistics

Accounting item	**Number of firm-year observations**	**Median (thousands euro)**	**Standard deviation (thousands euro)**	**Minimum value (thousands euro)**	**Maximum value (thousands euro)**
net profit	95 388	44 174	514 057	-31 298 361	32 224 897
total asset	95 811	996 078	7 876 588	0	309 664 000
sales	92 359	738 348	5 986 975	-3 143 696	372 513 433
equity	95 805	356 103	3 019 900	-3 055 733	217 127 070
Financial ratio	**Number of firm-year observations**	**Median (thousands euro)**	**Standard deviation (thousands euro)**	**Minimum value (thousands euro)**	**Maximum value (thousands euro)**
ROA	93 788	-0,20	15,36	-100	100
ROE	90 511	-3,79	64,24	-998,17	988,37
ROS	90 232	-465,54	92 799,77	-25 961 874	985 194

Source: Authors' own elaboration

The first objective of the study is to determine if datasets classified as the first category datasets are consistent with Benford's law. The second objective is to determine if datasets classified as the second category datasets, which are calculated on the basis of previous datasets are consistent with Benford's law.

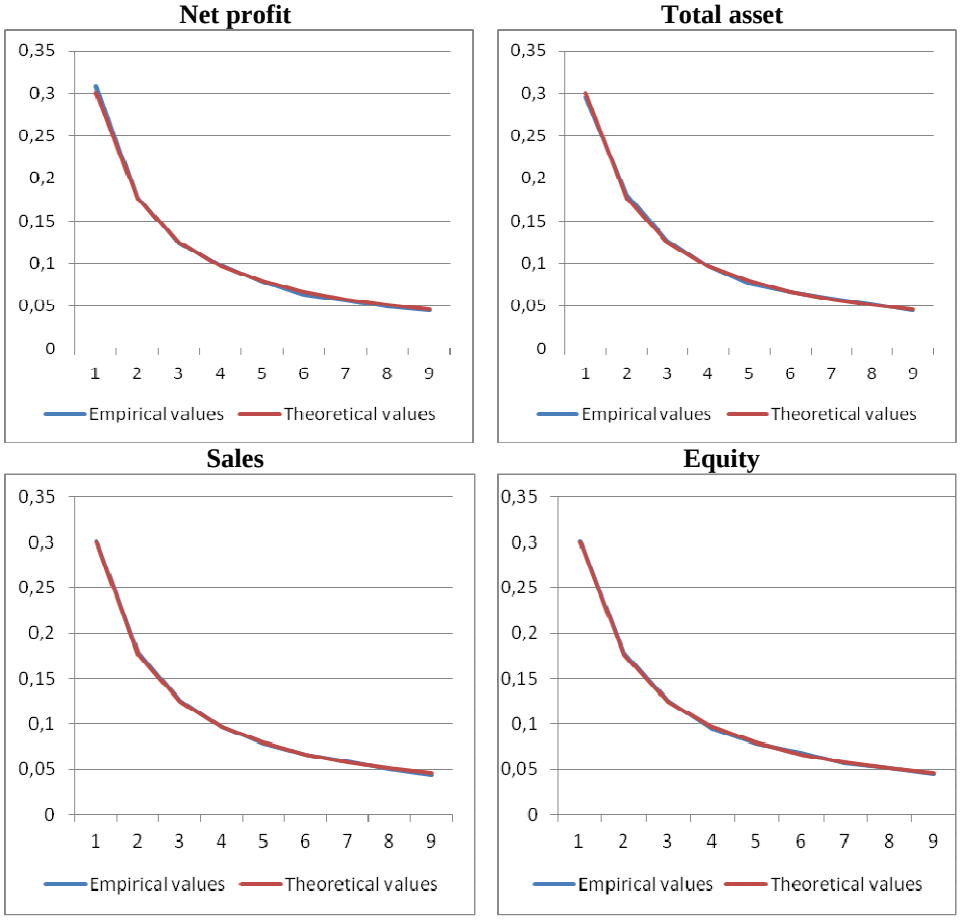
One of the properties of Benford's law, namely as arithmetic invariance allows to eliminate the negative values. Only absolute values are analyzed in the

study. Due to the fact, that many firm-year observations in the sample have very small values starting with 0-digit in the beginning of the number, all numbers in the datasets are multiplied by one thousand. In result in the majority of cases the first significant digit different from zero is taken into consideration in the study.

3. Analysis of the First Category Datasets

The first objective of the study is to analyze the distribution of the first significant leading digit in datasets representing net profit, total asset, sales and equity. Empirical values and theoretical values in the analyzed datasets are consistent with the distribution postulated by Benford's law (Fig. 1).

Fig.1 The Distribution of the First Leading Digit in Datasets Representing Accounting Items



Source: Authors' own elaboration

To examine the degree of adjustment of the distribution of the first significant leading digit with Benford's law we use two measures. The first measure is the absolute value of the difference between empirical and theoretical values of the distribution of the first leading digit. Theoretical values are calculated as the total number of observations in a given dataset multiplied by theoretical probability distribution proposed by Benford's law. We argue that this index represents the degree of how a given dataset comply with the distribution postulated by Benford's law.

In all analyzed accounting items the absolute value of the difference between empirical and theoretical values is less than two percent. The degree of adjustment is very high and we can conclude that analyzed datasets classified to the first category are consistent with Benford's law. It is worth mentioning that total sample sizes (Table 2) are slightly less numerous than in the initial sample (Table 1) because there are still some observations starting with zero-digit, despite the multiplication by a thousand. The second measure of adjustment – the absolute value of differences between empirical and theoretical number of observations also provides evidence of high degree of adjustment. In majority of cases the value of this measure is lower than 10%.

4. Analysis of the Second Category Datasets

Large datasets representing components of financial ratios like: net profit, total asset, sales and equity have distribution consistent with distribution postulated by Benford's law. At this stage of study, the question is if large datasets representing financial ratios have also distribution consistent with Benford's law. In order to focus analysis on the first significant digit, all numbers are multiplied by one thousand. In order to eliminate negative numbers only absolute values are taken into consideration. The distribution of the first leading digit is analyzed (Fig. 2).

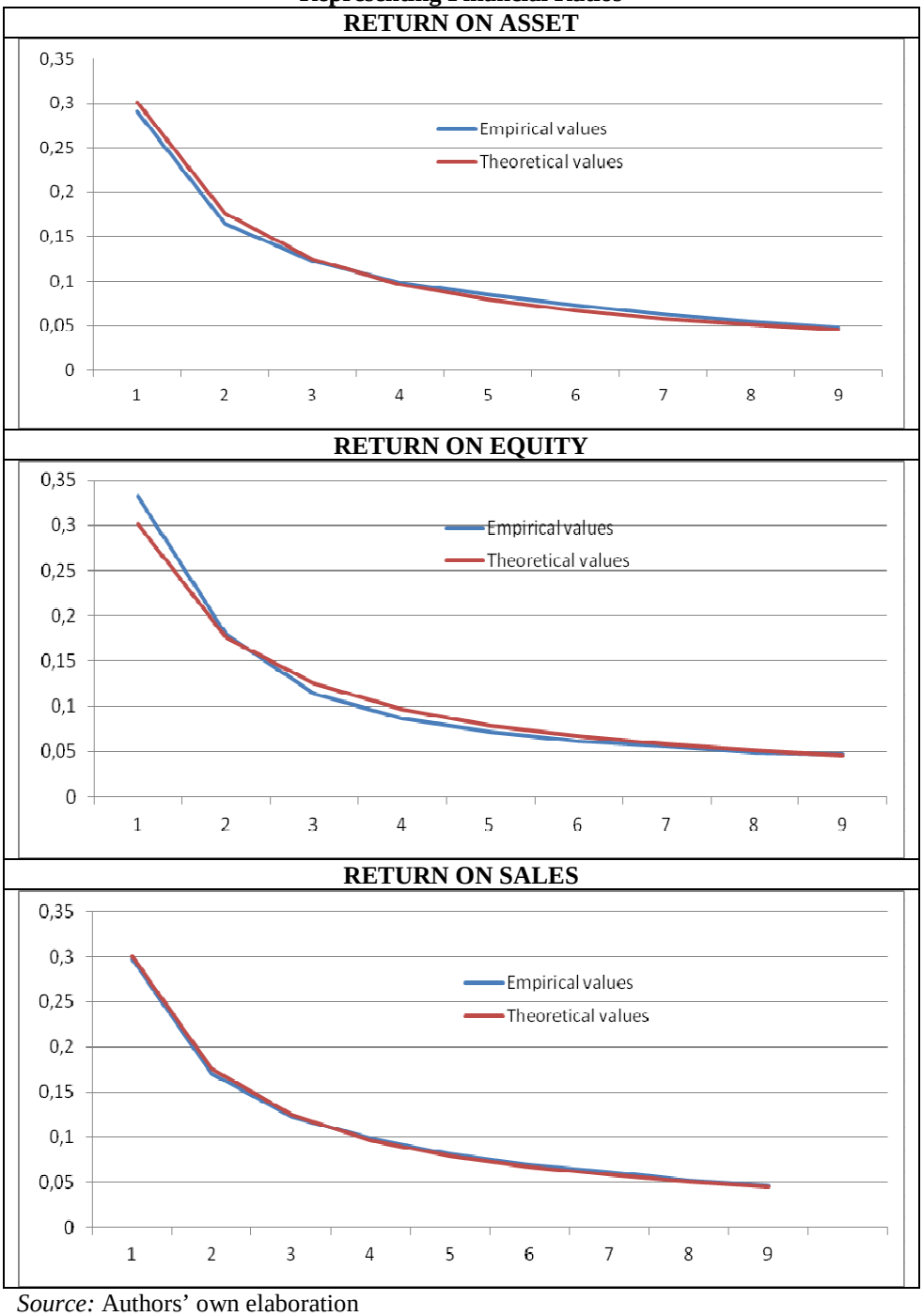
The difference between empirical and theoretical values is higher in large datasets representing financial ratios, than in datasets representing accounting items – components of financial ratios.

Table 2 Probability Distribution of the First Leading Digit in Datasets with Accounting Items

	First leading digit [1]	Number of observations [2]	Empirical values [3]	Theoretical values [4]	The difference [5] = [3] - [4]	The module difference [6] = |[5]|	Theoretical number of observations [7] = [4]*N	The module difference [8] = |([2] – [7])/ [7]
NET PROFIT	1	28 509	0,3080	0,3010	0,0070	0,0070	27 865	2,26%
	2	16 392	0,1771	0,1761	0,0010	0,0010	16 300	0,56%
	3	11 474	0,1240	0,1249	-0,0010	0,0010	11 565	0,79%
	4	9 014	0,0974	0,0969	0,0005	0,0005	8 971	0,48%
	5	7 265	0,0785	0,0792	-0,0007	0,0007	7 330	0,89%
	6	5 900	0,0637	0,0669	-0,0032	0,0032	6 197	5,03%
	7	5 258	0,0568	0,0580	-0,0012	0,0012	5 368	2,09%
	8	4 557	0,0492	0,0512	-0,0019	0,0019	4 735	3,91%
	9	4 198	0,0454	0,0458	-0,0004	0,0004	4 236	0,90%
	Sample size	**92 567**			**Total**	**0,0168**	**Total**	**16,91%**
TOTAL ASSET	**First leading digit [1]**	**Number of observations [2]**	**Empirical values [3]**	**Theoretical values [4]**	**The difference [5] = [3] - [4]**	**The module difference [6] = |[5]|**	**Theoretical number of observations [7] = [4]*N.**	**The module difference [8] = |([2] – [7])/ [7]**
	1	28 313	0,2956	0,3010	-0,0054	0,0054	28 835	1,84%
	2	17 323	0,1808	0,1761	0,0048	0,0048	16 867	2,63%
	3	12 107	0,1264	0,1249	0,0015	0,0015	11 968	1,15%
	4	9 296	0,0970	0,0969	0,0001	0,0001	9 283	0,14%
	5	7 328	0,0765	0,0792	-0,0027	0,0027	7 585	3,50%
	6	6 412	0,0669	0,0669	0,0000	0,0000	6 413	0,01%
	7	5 673	0,0592	0,0580	0,0012	0,0012	5 555	2,08%
	8	5 024	0,0524	0,0512	0,0013	0,0013	4 900	2,47%
	9	4 311	0,0450	0,0458	-0,0008	0,0008	4 383	1,67%
	Sample size	**95 787**			**Total**	**0,0178**	**Total**	**15,50%**
SALES	**First leading digit [1]**	**Number of observations [2]**	**Empirical values [3]**	**Theoretical values [4]**	**The difference [5] = [3] - [4]**	**The module difference [6] = |[5]|**	**Theoretical number of observations [7] = [4]*N.**	**The module difference [8] = |([2] – [7])/ [7]**
	1	27 169	0,3017	0,3010	0,0007	0,0007	27 109	0,22%
	2	16 068	0,1784	0,1761	0,0023	0,0023	15 858	1,31%
	3	11 320	0,1257	0,1249	0,0008	0,0008	11 251	0,61%
	4	8 695	0,0966	0,0969	-0,0004	0,0004	8 727	0,37%
	5	7 023	0,0780	0,0792	-0,0012	0,0012	7 131	1,53%
	6	6 022	0,0669	0,0669	-0,0001	0,0001	6 029	0,11%
	7	5 270	0,0585	0,0580	0,0005	0,0005	5 222	0,90%
	8	4 573	0,0508	0,0512	-0,0004	0,0004	4 606	0,73%
	9	3 913	0,0435	0,0458	-0,0023	0,0023	4 121	5,31%
	Sample size	**90 053**			**Total**	**0,0086**	**Total**	**15,50%**
EQUITY	**First leading digit [1]**	**Number of observations [2]**	**Empirical values [3]**	**Theoretical values [4]**	**The difference [5] = [3] - [4]**	**The module difference [6] = |[5]|**	**Theoretical number of observations [7] = [4]*N**	**The module difference [8] = |([2] – [7])/ [7]**
	1	28 472	0,3012	0,3010	0,0002	0,0002	28 455	0,06%
	2	16 922	0,1790	0,1761	0,0029	0,0029	16 645	1,64%
	3	11 855	0,1254	0,1249	0,0005	0,0005	11 810	0,38%
	4	8 965	0,0948	0,0969	-0,0021	0,0021	9 160	2,18%
	5	7 420	0,0785	0,0792	-0,0007	0,0007	7 485	0,87%
	6	6 451	0,0682	0,0669	0,0013	0,0013	6 328	1,91%
	7	5 367	0,0568	0,0580	-0,0012	0,0012	5 482	2,14%
	8	4 832	0,0511	0,0512	0,0000	0,0000	4 835	0,07%
	9	4 240	0,0449	0,0458	-0,0009	0,0009	4 325	2,01%
	Sample size	**94 524**			**Total**	**0,0098**	**Total**	**11,24%**

Source: Authors' own elaboration

Fig. 2 The Distribution of the First Leading Digit in Datasets Representing Financial Ratios



Source: Authors' own elaboration

Table 3 Probability Distribution of the First Leading Digit in Financial Ratios

	First leading digit [1]	Number of observations [2]	Empirical values [3]	Theoretical values [4]	The difference [5] = [3] - [4]	The module difference [6] = |[5]|	Theoretical number of observations [7] = [4]*N	The module difference [8] = |([2] – [7])/ [7]
RETURN ON ASSET	1	27 091	0,2902	0,3010	-0,0109	0,0109	28 104	3,74%
	2	15 354	0,1645	0,1761	-0,0116	0,0116	16 440	7,07%
	3	11 454	0,1227	0,1249	-0,0023	0,0023	11 664	1,84%
	4	9 242	0,0990	0,0969	0,0021	0,0021	9 048	2,10%
	5	7 910	0,0847	0,0792	0,0055	0,0055	7 392	6,54%
	6	6 841	0,0733	0,0669	0,0063	0,0063	6 250	8,64%
	7	5 826	0,0624	0,0580	0,0044	0,0044	5 414	7,07%
	8	5 113	0,0548	0,0512	0,0036	0,0036	4 776	6,60%
	9	4 529	0,0485	0,0458	0,0028	0,0028	4 272	5,68%
	Sample size	**93 360**			**Total**	**0,0495**	**Total**	**49,28%**
	First leading digit [1]	**Number of observations [2]**	**Empirical values [3]**	**Theoretical values [4]**	**The difference [5] = [3] - [4]**	**The module difference [6] = |[5]|**	**Theoretical number of observations [7] = [4]*N**	**The module difference [8] = |([2] – [7])/ [7]**
RETURN ON EQUITY	1	29 949	0,3323	0,3010	0,0313	0,0313	27 131	9,41%
	2	16 208	0,1798	0,1761	0,0037	0,0037	15 871	2,08%
	3	10 299	0,1143	0,1249	-0,0107	0,0107	11 260	9,33%
	4	7 822	0,0868	0,0969	-0,0101	0,0101	8 734	11,66%
	5	6 503	0,0722	0,0792	-0,0070	0,0070	7 136	9,74%
	6	5 569	0,0618	0,0669	-0,0052	0,0052	6 034	8,34%
	7	5 040	0,0559	0,0580	-0,0021	0,0021	5 227	3,70%
	8	4 488	0,0498	0,0512	-0,0014	0,0014	4 610	2,72%
	9	4 249	0,0471	0,0458	0,0014	0,0014	4 124	2,94%
	Sample size	**90 127**			**Total**	**0,0728**	**Total**	**59,94%**
	First leading digit [1]	**Number of observations [2]**	**Empirical values [3]**	**Theoretical values [4]**	**The difference [5] = [3] - [4]**	**The module difference [6] = |[5]|**	**Theoretical number of observations [7] = [4]*N**	**The module difference [8] = |([2] – [7])/ [7]**
RETURN ON SALES	1	26 097	0,2969	0,3010	-0,0041	0,0041	26 458	1,38%
	2	14 985	0,1705	0,1761	-0,0056	0,0056	15 477	3,28%
	3	10 798	0,1229	0,1249	-0,0021	0,0021	10 981	1,69%
	4	8 677	0,0987	0,0969	0,0018	0,0018	8 518	1,84%
	5	7 190	0,0818	0,0792	0,0026	0,0026	6 959	3,21%
	6	6 055	0,0689	0,0669	0,0019	0,0019	5 884	2,82%
	7	5 413	0,0616	0,0580	0,0036	0,0036	5 097	5,84%
	8	4 582	0,0521	0,0512	0,0010	0,0010	4 496	1,88%
	9	4 094	0,0466	0,0458	0,0008	0,0008	4 022	1,77%
	Sample size	**87 891**			**Total**	**0,0236**	**Total**	**23,71%**

Source: Authors' own elaboration

In all analyzed profitability ratios, the absolute value of the difference between empirical and theoretical probability is much higher than in accounting items, ranging from 2,36% (in the case of return on sales) up to 7,28% (in the case of return on equity). The degree of adjustment is not as high as for previously analyzed accounting items, however, it is on acceptable level for datasets representing return on sales and return on equity. The second measure of adjustment – the absolute value of differences between empirical and

theoretical numbers of observations, shows lower level of adjustment, but still in majority of first leading digits the difference is lower than 10%.

Again, total sample sizes (Table 3) are slightly less numerous than in the initial sample (Table 1), because there are still some observations starting with zero-digit.

5. Conclusions

The main objective of the analysis is to examine compliance of large financial datasets with Benford's law. Two stage analysis starts with examining datasets representing accounting items: net profit, total assets, equity and sales. The results provide evidence, that the distribution of the first leading digit is consistent with the distribution postulated by Benford's law. In the second stage of the analysis, datasets representing profitability ratios: return in asset, return on equity and return on sales have been investigated.

Table 4 Summary of the Results

Measure of adjustement	Total assets	Net profit	Sales	Equity	Return on assets	Return on equity	Return on sales
Total sum of absolute differences between probability of theoretical and empirical distribution of the first leading digit	0,0178	0,0168	0,0086	0,0098	0,0495	0,0728	0,0236
Total sum of absolute differences between theoretical and empirical numbers of observations.	15,5%	16,9%	11,1%	11,2%	49,3%	59,9%	23,7%

Source: Authors' own elaboration

Based on the performed study, it can be concluded that large financial datasets representing accounting items reported by publicly listed companies in Europe are consistent with Benford's law. To a somewhat lesser extent, Benford's law is valid to large datasets representing financial ratios. Assuming Benford's law as a reliability criterion, it can be concluded that all analyzed datasets are reliable.

References

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

Bhattacharya, S., Dongming, X., Kumar, K. (2010) An ANN-based auditor decision support system using Benford's law. *Decision Support Systems*, 50 (3): 576-583.

Breunig, Ch., Goerres, A. (2011) Searching for electoral irregularities in an established democracy: Applying Benford's Law tests to Bundestag elections in Unified Germany. *Electoral Studies*, 30 (3): 534-545.

Lobert, T. (2008) On the non-existence of a general Benford's Law. *Mathematical Social Science*, 55, 103-106.

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

Orita M., Moritomo A., Niimi T., Ohno K. (2010) Use of Benford's law in drug discovery data. *Drug Discovery Today*, 15, 328-331.

ISPITIVANJE POUZDANOSTI VELIKIH FINANSIJSKIH SKUPOVA PODATAKA KORIŠĆENJEM BENFORDOVOG ZAKONA

Apstrakt: Autori su u radu analizirali velike skupove finansijskih podataka iz perspektive usklađenosti sa Benfordovim zakonom. Istraživane su dve grupe skupova podataka: izabrane računovodstvene pozicije iskazane od strane javno kotiranih evropskih korporacija i indikatori profitabilnosti. Smatramo da ako su skupovi podataka koji predstavljaju komponente finansijskih pokazatelja pouzdani iz perspektive Benfordovog zakona da su skupovi podataka koji predstavljaju finansijski pokazatelji zasnovani na ovim računovodstvenim pozicijama takođe pouzdani. Predstavljeno istraživanje pruža dokaze da su izvedeni skupovi podataka visokog kvaliteta, pod uslovom da su izvorni veliki skupovi finansijskih podataka visokog kvaliteta.

Ključne reči: Benfordov zakon, koeficijent profitabilnosti, pouzdanost finansijskih podataka