Novi Ekonomist
Vol 19(2), Year XIX, Issue 38, july – december 2025.
ISSN 1840-2313 (Print) 2566-333X (Online)
DOI: 10.69781/NOEK202538170

Submitted: 11.04.2025.
Accepted: 15.05.2025.
ORIGINAL ARTICLE
UDK: 336.021:[657.632:005.915

USING AI TO VERIFY AND ANALYZE BENFORD'S LAW IN REAL DATA

Vesna Rajić
Faculty of Economics, University of Belgrade, Serbia
vesna.rajic@ekof.bg.ac.rs
ORCID: 0000-0002-4566-0147

Jelena Stanojević
Faculty of Economics, University of Belgrade, Serbia
jelena.stanojevic@ekof.bg.ac.rs
ORCID: 0000-0002-4566-0147

Paper presented at the 13th International Conference EKONBIZ 2025 - Economic implications of the multipolar world, June 5th and 6th 2025, Bijeljina.

Abstract: *Benford's law is a key tool for detecting irregularities and potential manipulations in numerical data sets. This law describes the probability of the appearance of the first digits in large sets of numerical values, which allows for the identification of anomalies and verification of authenticity in them. The subject of this paper is the application of artificial intelligence in the analysis and verification of Benford's law on real data. Given the increasingly widespread application of artificial intelligence in the automation of data analysis, fraud detection and statistical verification of economic and financial reports, the aim of the paper is to explore the possibilities of using machine learning algorithms, such as deep neural networks and classification methods, to recognize and analyze deviations from the expected distribution of the first digits. The use of artificial intelligence in the automation of the verification process and the detection of manipulations in data is also considered. The results show that the application of artificial intelligence can significantly improve the accuracy of anomaly detection, while at the same time enabling faster and more efficient analysis of large data sets. It is concluded that artificial intelligence is a powerful tool in improving the application of Benford's law in practical situations, especially in the analysis of financial and other types of data.*

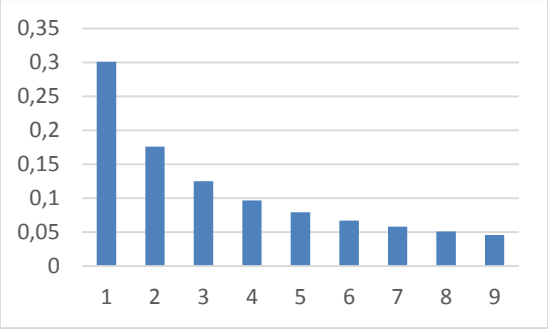
Key words: *Benford's law, AI, simulations, manipulation*

JEL classification*: C45, C46, C63*

1. INTRODUCTION

Benford's law is a statistical law and phenomenon that describes the distribution of the first digit in many real-world data sets. According to this law, smaller numbers (such as 1 and 2) appear as the first digit significantly more often than larger numbers (1 occurs 30.1% of the time, 2 occurs 17.6% of the time, and 9 occurs 4.6%), which is contrary to the intuitive expectation of a uniform distribution. The following figure illustrates Benford's law using a probability diagram.

Figure 1. Benford's distribution



Source: Output from Excel

This law is widely used in fraud detection, forensic accounting analysis, analysis of economic and financial reports, tax investigations, analysis of experimentally obtained data in various scientific fields, demographic statistics, analysis of election results and other areas where it is necessary to verify the authenticity of data.

The development of artificial intelligence (AI) has opened up new opportunities for automated verification and analysis of Benford's law in large data sets. Using machine learning algorithms and deep data analysis methods, AI can effectively identify deviations from the expected distribution and indicate potential irregularities.

The aim of this paper is to explore how artificial intelligence can be applied in the analysis of real data in the context of Benford's law. The focus will be on the application of various AI techniques and algorithms in detecting deviations and identifying data that do not follow this statistical regularity. We will also consider their application in automating the verification process and detecting data manipulation. Attention will also be paid to analyzing the advantages and challenges of this approach, as well as its practical application in various areas.

2. BENFORD'S LAW AND ARTIFICIAL INTELLIGENCE: LITERATURE REVIEW

Benford's law was first observed by Simon Newcombe in 1881, while its first formulation was given by Frank Benford in 1938. In his work "The Law of Anomalous Numbers", he analyzed 20,000 data from various sources and observed that the first digit of a number with at least four digits appears according to a decreasing logarithmic law. Namely, it was observed that the digit 1 appears with the highest frequency, about 30%, while the digit 9 appears with the lowest frequency of 4.58%.

There is extensive literature dealing with Benford's law. It includes theoretical foundations, applications and modern approaches to data analysis. The mathematical framework of the law was established in the early works, while more recent research has focused mainly on its application in auditing, fraud detection, and data analysis with the help of artificial intelligence. Below, we list the sources that represent key resources for researching Benford's law, both in theory and in practice.

Hill (1995) provided a mathematical analysis of Benford's law, using probability theory to explain why data from different sources often follow this distribution. Pinkham (1961) proved that Benford's law is invariant under changes in units, which further confirms its universality in real data. Also, for many data sets, such as natural numbers, economic data, city populations, or financial data, this rule does not change, and deviations from this distribution may indicate irregularities or manipulation of the data. Nigrini (1996) investigated the application of Benford's law in detecting tax fraud and financial data manipulation. Durtschi et al. (2004) analyzed how Benford's law is used in accounting investigations to detect fraud, providing concrete examples from practice. A book that discusses in detail the methods of applying Benford's law in forensic accounting and auditing, with case studies and algorithms for fraud detection is Nigrini (2012). Furthermore, for the basic literature dealing with machine learning techniques, as one of the basic methods of artificial intelligence, we highlight Alpaydin (2020). We highlight the following authors, who have dealt with Benford's law and artificial intelligence. Varian (1972) uses advanced methods in his work to examine the statistical characteristics of Benford's law and the possibilities of its application. Kossovsky (2014) provides a detailed mathematical and practical analysis of Benford's law in his book, with the application of machine learning methods in data analysis. Bishop & Nasrabadi (2006) provide a basic structure and explanation of how to use AI to analyze and verify Benford's law in their work.

3. BENFORD'S DISTRIBUTION

In this section, we give basic definitions and theorems, which are the first step for further research on Benford's law.

Definition 1: For every real number x, there exists an integer part $\lfloor x \rfloor$ and a fractional part $\langle x \rangle$, such that x can be expressed uniquely: $x = \lfloor x \rfloor + \langle x \rangle$, where $\lfloor x \rfloor = \max\{k \in Z: k \leq x\}$ and $\langle x \rangle = x - \lfloor x \rfloor$.

Definition 2: (Significance) For every positive number $x > 0$ and base B, x can be represented in the following form: $x = S_B(x) \cdot B^{k(x)}$, where $S_B(x) \in [1, B)$ is the significance of x and the integer $k(x)$ (necessarily unique) is the exponent. For a negative number x: $S_B(x) = S_B(-x)$ and $S_B(0) = 0$.

Definition 3: (Benford's law) A random variable X obeys Benford's law in the basis B if $\forall s \in [1, B)$,

$$P(S_B(x) \leq s) = log_B(s),$$

in particular,

$$P(D_1(X) = x) = log_B\left(1 + \frac{1}{x}\right),$$

where $x \in \{1,2, \dots, B-1\}$ is a digit on the first position in the number that takes value in the set of all possible outcomes and $D_1(X)$ is a random variable representing the first left digit of $S_B(x)$, that is first significante digit of X.

If the data follows Benford's law, then we can say that they have a Benford's law distribution. For a random variable X, which represents the first digit, specifically in base $B = 10$, the following holds:

$P(X = x) = log_{10}\left(1 + \frac{1}{x}\right)$,
$x = 1, 2, \dots, 9$, while the formula for the distribution law function is:

$F(x) = P(X \leq x) = log_{10}(1 + x)$.

Next, we give the formulations of two fundamental theorems, without proof. The first result tells how a random variable can be generated that has a Benford's distribution, while the second result gives a necessary and sufficient condition for a random variable to have a Benford's distribution.

Theorem 1: (Berger & Hill, 2021) If $X \sim Uniform(0,1)$ then $Y = 10^X$ conforms to Benford's law.

Theorem 2: : (Fang & Chen, 2020) A random variable $X > 0$, follows Benford's law in base B iff the random variable $Y = \langle log_B(X) \rangle$, the fraction part of $log_B(X)$ is uniformly distributed in [0,1].

4. ARTIFICIAL INTELLIGENCE AND ITS APPLICATION IN DATA ANALYSIS

The revolutionary technology, whose dynamic development we are currently observing, is completely changing the process of manipulating data, more precisely: the way in which it is collected, processed and analyzed. Organizations and entities that apply artificial intelligence in data analysis are able to make more reliable decisions, improve their efficiency and business, as well as discover patterns of behavior in the collected and analyzed data, which would most likely remain unnoticed. Here we will consider how artificial intelligence can improve data analysis, what methods are used for this purpose and what challenges arise in its application.

In order to automate data analysis and maximize the extraction and use of useful information, artificial intelligence uses various methods and techniques. Some of the methods that stand out for their importance and application are: machine learning, deep learning, natural language processing, and computer vision. All of them allow computer systems to learn from available data and, over time, improve their performance. Since artificial intelligence enables automatic analysis of data and recognition of patterns in it, which the human eye would most likely miss, it is particularly used in the analysis of large data sets.

Machine learning is unique because its algorithms allow systems to learn from data and make decisions or predictions without having to write or use explicit programs or programming. Deep learning is a subfield of machine learning that uses neural networks with multiple layers to recognize complex patterns in large data sets. In particular, in text analysis, natural language processing techniques allow computers to understand, interpret, and generate human language. Systems are also able to analyze and understand visual data, which is essential for image and video analysis, using computer vision, which sets it apart from all known and available methods. Some of the most significant examples of the application of artificial intelligence in various areas of data analysis are as follows:

1. Predictive analytics - Historical data is used to predict trends and user behavior.

2. Fraud detection - Artificial intelligence can be used to analyze large amounts of transactional data to identify suspicious activity and prevent possible fraudulent actions.

3. User experience personalization - Algorithms for providing personalized content to users are already used in social networks, e-commerce, and internet streaming, which is multimedia content in which the user himself simultaneously receives and plays the content, unlike classic download methods, where the user must wait for the download to complete in order to play the content.

4. Health data analysis - Artificial intelligence can help diagnose diseases, predict epidemics, and improve health treatments, all of which can improve the level of healthcare in a society.

5. Business process optimization - Companies large and small can improve supply chain efficiency and decision-making processes, as well as reduce business costs, by using artificial intelligence.

5. APPLICATION OF ARTIFICIAL INTELLIGENCE IN VERIFICATION OF BENFORD'S LAW

In this section, we outline the key processes through which artificial intelligence can be used to verify Benford's law, along with an explanation of specific methods for applying artificial intelligence in data analysis based on Benford's law.

Specifically, in the context of Benford's law analysis, artificial intelligence can be used in the following steps and processes:

1. Pattern recognition - Recognizing patterns in data that agree with Benford's law and also identifying deviations from the law, with high accuracy and precision, can be detected by machine learning algorithms.

2. Anomaly detection - Artificial intelligence is able to identify anomalies in data, which may indicate manipulations and errors, both accidental and intentional, which are of significant interest to identify and detect.

3. Automation of data verification - Artificial intelligence is able to automatically check whether data conforms to the distribution predicted by Benford's law. This allows its application in many situations and analyses, such as, for example, in forensic analysis of financial statements or data validation in scientific and professional research.

Next, we provide an overview of specific methods of how artificial intelligence can be applied in data analysis according to Benford's law. First, data classification - data classification algorithms, such as support vector machines, k-nearest neighbors, or neural networks, can be used to classify data according to their compliance with Benford's law. Also, these algorithms can learn the rule of distribution of the first digits and further use it to evaluate new data sets. Second, anomaly detection algorithms - since artificial intelligence is able to automatically analyze large data sets and identify suspicious deviations from the Benford's distribution, their importance and implementation are clear. These algorithms, such as "random forests" or "deep network anomaly detection", are able to identify data that does not correspond to the expected distribution, which may indicate errors, fraud or manipulation.

Here we will also highlight algorithms for data simulation and verification - artificial intelligence can also be used to generate simulations, and thus it is possible to create models that generate data according to the Benford's distribution, and then compare the real data with the simulated ones, in order to determine whether the data is in accordance with the law.

The application of artificial intelligence for data analysis and manipulation, which includes the verification and application of Benford's law, is quite broad. Some examples are:

- financial statements, whose analysis can be used by artificial intelligence to detect potential irregularities, such as falsely reported income or manipulations in the reports of a particular company;

- statistical data and population, where artificial intelligence is able to detect anomalies in the distribution of numbers, when analyzing population data, which can indicate errors in censuses or manipulated statistics;

- scientific data, which are obtained from experiments in scientific research, artificial intelligence can analyze and determine whether they behave in accordance with Benford's law;

- digital marketing where AI is used to improve the personalization, optimization and analysis of marketing campaigns in order to check behavior in accordance with certain patterns (about digital marketing see for example Mihailović et al., 2024);

- risk management where AI is used for faster detection of unforeseen situations, better decision-making, as well as proactive response to potential problems (see Arshi, 2022).

We can certainly expect AI to become an even more integral tool in all business areas, providing organizations with competitive advantages in data analysis, financial planning and marketing strategies.

6. DATA AND METHODOLOGY

Due to the challenges in the healthcare system caused by the COVID-19 pandemic, the consequences of which are still being felt today, we decided to focus in this paper on the area of health insurance. Namely, as the COVID-19 pandemic led to a significant increase in healthcare costs, this imposed the need for increased supervision and control, especially with regard to financial reports. Benford's law, which is commonly used to identify potential anomalies in data sets, was applied to examine irregularities in the financial reports of three private hospitals in the Republic of Serbia, in the paper Stanojević et al. (2024). We will use data for one hospital (which has already been analyzed in the aforementioned paper) and check compliance with Benford's law using artificial intelligence. The data are certainly available on the website of the Serbian Business Registers Agency [1] .

The balance sheet is a financial statement that provides an overview of a hospital's assets, as well as its liabilities, or sources of financing. On the other hand, the income statement shows the hospital's revenues and expenses, allowing the determination of the business results (profit or loss) over a certain period of time, Meiryani et al. (2020). Since financial statements can be subject to manipulation and fraud by reporters in order to obtain additional subsidies or financial incentives, especially in times of crisis, such as the COVID-19 pandemic, we discuss this aspect here. Below, we will show how we check for compliance with Benford's law, using artificial intelligence (step by step).

1. *Extraction of the first digits*: For each number in the analyzed data set, we will extract the first digit (for example, for the number 986871, the first digit is 9).

2. *Calculating the distribution of the first digits*: We will calculate how often each digit from 1 to 9 appears in considered data set.

[1] www.apr.gov.rs

3. *Comparison with the theoretical distribution according to Benford's law*: Benford's law predicts the following probability for the occurrence of the first digit d (where d is a number from 1 to 9):

$$P(X = d) = log_{10}\left(1 + \frac{1}{d}\right).$$

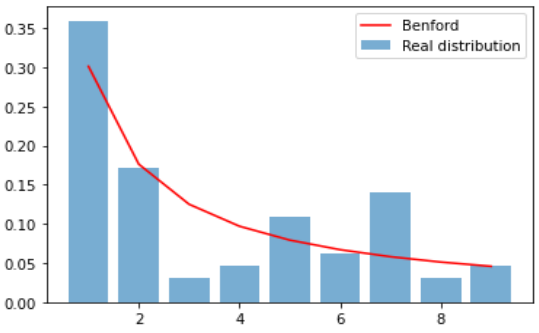
4. *Statistical test (Chi-square test)*: We will use the Chi-square test to check whether the actual distribution is significantly different from the theoretical distribution according to Benford's law.

5. *Visualization*: We will create a graph to better see the differences.

If we want to use advanced AI tools, we can use specialized libraries and platforms. For example, we can use Python libraries: benford, numpy, pandas, scipy and matplotlib. The Python code for the analysis is:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import chisquare

# Data
data = [....]

# Function to extract the first digit
def first_digit(number):
    return int(str(number)[0])

# Drawing the first digit
first_digits = [first_digit(num) for num in data]

# Distribution of the first digits
digit_counts=
pd.Series(first_digits).value_counts().sort_index()

# Theoretical distribution according to the
Benford's law
benford_dist = [np.log10(1 + 1/d) for d in range(1,
10)]

# Visualization
plt.bar(range(1, 10), digit_counts /
sum(digit_counts), alpha=0.6, label='Real
distribution')
plt.plot(range(1, 10), benford_dist, 'r',
label='Benfordov zakon')
plt.legend()
plt.show()

# Chi-square test
chi2_stat, p_val = chisquare(digit_counts,
f_exp=np.array(benford_dist) * sum(digit_counts))
print(f'Chi-square stat: {chi2_stat}, p-value:
{p_val}')
```

Figure 2. Checking compliance with the Benford's law



Chi-square stat: 15.6686011589155, p-value: 0.04737665421212908

Source: Python output

The lines in this code are as follows:

1. The first function (first_digit) extracts the first digit from each number.

2. The distribution of the first digits is calculated by counting the number of times each digit appears.

3. Benford's distribution is calculated according to the formula: $log_{10}\left(1 + \frac{1}{d}\right)$ for d from 1 to 9.

4. The visualization compares the empirical distribution with the theoretical distribution.

5. The Chi-square test is used to check whether there are statistically significant differences between the actual and theoretical distributions. The Chi-square statistic is calculated as:

$$\chi^2 = \sum_{i=1}^{9} \frac{(O_i - E_i)^2}{E_i} = n \sum_{i=1}^{9} \frac{(p_{oi} - p_i)^2}{p_i} ,$$

where O_i is the observed frequency of the digit i, E_i is the expected frequency of the digit i according to the Benford's law.

7. RESULTS

Running the code in Python we got the following results:

• Chi-square stat: 15.67

• p-value: 0.047

Since the p-value is greater than 0.01, we conclude that we do not have enough evidence to claim that the analyzed data significantly differ from Benford's law. This means that at the 0.01 significance level we cannot reject the hypothesis that the data follows Benford's law. In other words, the analyzed data can follow Benford's law with a high degree of certainty (at the 1% level).

CONCLUSION

Based on the analysis of data from the financial statements of a private hospital, Benford's law was applied to check the consistency with the theoretical distribution of the first digits. The analysis used data on the hospital's income and expenses, and then artificial intelligence methods and statistical tests were applied to determine whether there is a statistically significant difference between the actual and theoretical distribution of the first digits. Based on the results obtained, we can conclude that we do not have enough evidence to claim that the data significantly differs from the theoretical distribution of the first digits according to Benford's law. This analysis is an example of how artificial intelligence can be used in practice to recognize patterns and detect potential irregularities in financial data. By using advanced AI techniques such as statistical tests, pattern recognition algorithms and processing of large data sets, we can effectively spot possible anomalies in financial reporting, which is crucial in the context of increased control and surveillance in crisis periods, such as the COVID-19 pandemic. In the future, AI may become an even more important tool for automatically detecting potential fraud and ensuring greater transparency in the financial sector. Future research directions relate to the application of AI in various economic disciplines, because AI is a key technology that is increasingly shaping various aspects of the business world, from data analysis and financial reports to digital marketing and risk management. In the context of financial reporting, AI offers significant advantages in detecting irregularities and potential fraud.

REFERENCES

[1] Alpaydin, E. (2020). Introduction to machine learning. MIT press, Cambridge.

[2] Arshi, N. (2022). Role of Artificial Intelligence in business risk management MANAGEMENT. American Journal of Business Management, Economics and Banking, 1, 55–66.

[3] Benford, F. (1938). The law of anomalous numbers. Proceedings of the American philosophical society, 78(4), 551-572.

[4] Berger, A., Hill, T. P. (2021). The mathematics of Benford's law: a primer. Statistical Methods & Applications, 30(3), 779-795.

[5] Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning, 4(4), p. 738, New York: springer.

[6] Costa, J., Travassos, S., Santos, J. (2013, June). Application of Newcomb-Benford law in accounting audit: a bibliometric analysis in the period from 1988 to 2011. In 10th International Conference on Information Systems and Technology Management-CONTECSI.

[7] Durtschi, C., Hillison, W., Pacini, C. (2004). The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data. Journal of Forensic Accounting, 5(1), 17-34.

[8] Fang, G., & Chen, Q. (2020). Several common probability distributions obey Benford's law. Physica A: Statistical Mechanics and its Applications540, https: //doi.org /10.1016 /j.physa. 2019.123129.

[9] Hill, T. P. (1995). Base-invariance implies Benford's law. Proceedings of the American Mathematical Society, 123(3), 887-895.

[10] Kossovsky, A. E. (2014). Benford's law: theory, the general law of relative quantities, and forensic fraud detection applications (Vol. 3). World Scientific, https://doi.org/10.1142/9089

[11] Meiryani, M., Soepriyanto, G., D. Wahyuningtias, D. & Dewi, K. (2020). Accounting Perspective in Hospital. International Journal of Online and Biomedical Engineering, 16(8), 114-123.

[12] Mihailović, B. M., Radosavljević, K., Popović, V., & Puškarić, A. (2024). Impact of digital marketing on the performance of companies in the agricultural sector of Serbia. Economics of Agriculture, 71(1), 173-188.

[13] Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. American Journal of mathematics, 4(1), 39-40.

[14] Nigrini, M. J. (1996). A taxpayer compliance application of Benford's law. The Journal of the American Taxation Association, 18(1), 72-91.

[15] Nigrini, M. J. (2012). Benford's Law: Applications for forensic accounting, auditing, and fraud detection (Vol. 586). John Wiley & Sons.

[16] Pinkham, R. S. (1961). On the Distribution of First Significant Digits. The Annals of Mathematical Statistics, 32(4), 1223-1230.

[17] Stanojević, J., Radojičić, D., Rajić, V., & Rakonjac-Antić, T. (2024). Statistical analysis of fitting Pareto and Weibull distributions with Benford's Law: theoretical approach and empirical evidence. Hacettepe Journal of Mathematics and Statistics, 53(6), 1724-1741.

[18] Varian, H. R. (1972). Benfords law. American Statistician, 26(3), 65-66.