Novi Ekonomist
Vol 18(1), Year XVIII, Issue 35, january – june 2024.
ISSN 1840-2313 (Print) 2566-333X (Online)
DOI: 10.69781/NOE202435014

Submitted: 19.04.2024.
Accepted: 26.04.2024.
Scientific article
UDK: 330.322:336.748.12(100)

AN OVERVIEW OF THE DISTRIBUTIONS OF CLAIMS USED IN NON-LIFE INSURANCE THAT CONFORM TO BENFORD'S LAW

Jelena Stanojević
University of Belgrade, Faculty of Economics and Business, Belgrade, Serbia
jelena.stanojevic@ekof.bg.ac.rs
ORCID: 0000-0001-5668-5297

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

Paper presented at the 12th International Scientific Symposium „EkonBiz" - New Economic Reality: The Economic Consequences of Social and Demographic Transition, Bijeljina, 30th and 31st May 2024.

Abstract: *Benford's law, known as the law of the first digit, is used as a basic method to identify possible manipulation, or errors detection, in a large data set. Namely, according to that law, the first digits of the corresponding data set appear with a frequency determined according to the decreasing logarithmic law, which is contrary to the intuition about their uniform appearance. Thus, the number 1 appears in approximately 30% of the cases as the first digit, while the number 9 appears in 4.58% of cases. The subject of the paper is an overview of distributions that conform to Benford's law, which is confirmed both theoretically (by proving the theorems) and empirically (by conducting simulations). The goal of the paper is to determine the distribution that best fits the data on non-life insurance claims and then to examine the agreement with Benford's law using statistical tests applied in the literature. The main result of the paper is to determine the agreement or disagreement of the observed data set with Benford's law, thus providing an answer to the question about their possible manipulation and a possible proposal for a deeper analysis of individual numbers.*

Key words: *Benford's law, manipulation, Benford's law tests, simulations, non-life insurance*

JEL classification: *C46, G22*

1. INTRODUCTION

Benford's law is named after the scientist Frank Benford, who in 1938 published a paper entitled "The Law of Anomalous Numbers" in the journal *Proceedings of the American Philosophical Society*. In his paper, he analyzed 20,000 data from different sources and concluded "that there is a logarithmic distribution of first digits when the numbers are composed of four or more digits". Before him, Simon Newcomb dealt with this problem in 1881. Actually, he was the first to notice that the first digits of the corresponding data set appear with a frequency determined according to the decreasing logarithmic law.

According to Benford's law, the occurrence of the first digits is not uniform. Actually, it claims that many numerical data sets follow the trend that leading digits 1-9 appear with decreasing logarithmic distribution, where digit 1 appears with the largest frequency, almost 30%, and digit 9 with a frequency of 4.58%.

There is a vast literature which covers this topic. We will mention only a few sources. For example, Warshavsky (2010) presented application of the law in the area of accounting. Nigrini (1996) wrote about the law in the detection of fraud, in the area of accounting, auditing, and taxation. Further, Hill (1995) suggested a strict proof of the law based on mathematical theory.

As the data on reported losses in property insurance are subject to some kind of manipulation and fraud, as well as random errors (for example, when entering data, recording results or necessary rounding), there is a need to apply a methodology that will locate suspect numbers. In this paper, we analyzed such data on the values of losses during the period 1997-2002 in the tariff fire-industry. The data was obtained from one insurance company for one city in Serbia. First, a distribution that fits the data was determined. Two distributions passed the goodness of fit test, and both of them conform to Benford's law which is shown in the literature. Furthermore, four tests were applied in order to check the agreement of the appearance of the first digit of the observed data with Benford's law, and the corresponding conclusions were given.

2. BENFORD’S LAW AND BENFORD’S DISTRIBUTION

We give some standard definitions and theorems.

Definition 1: For every real number x there exists the integer part $\lfloor x \rfloor$ and the fractional part $\langle x \rangle$ of x, where x can be expressed uniquely as $x = \lfloor x \rfloor + \langle x \rangle$. It is satisfied that $\lfloor x \rfloor = \max\{\mathrm{k} \in \mathrm{Z}: \mathrm{k} \leq \mathrm{x}\}$ and $\langle x \rangle = x - \lfloor x \rfloor$.

Definition 2: (Significand) For any positive number $x > 0$ and base B, x is represented in notation as $x = S_B(x) \cdot B^{k(x)}$, where $S_B(x) \in [1, B)$ is called the significand of x and the integer $k(x)$ (necessarily unique) represents the exponent. For negative number x it is satisfied $S_B(x) = S_B(-x)$ and $S_B(0) = 0$.

Definition 3: (Benford's law) A random variable X obeys Benford's law in base B if for any $s \in [1, B)$,

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

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 data obeys Benford’s law, we can also say that it follows Benford distribution. If we consider random variable X as a first digit, the probability mass function of Benford’s distribution, for base $B = 10$, is

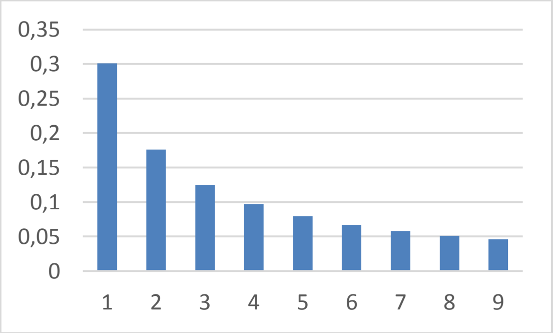
$$P(X = x) = log_{10}\left(1 + \frac{1}{x}\right),$$

$x = 1, 2, \dots, 9$, while the cumulative distribution function (cdf) is:

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

We can represent the Benford distribution graphically using a probability diagram. This diagram is shown in Figure 1.

***Figure 1.** Benford’s distribution*



***Source:** Output from Excel*

Some more features of this distribution are as follows:

-the expected value is equal to:

$$E(X) = \sum_{x=1}^{9} xlog_{10}\left(1 + \frac{1}{x}\right) \approx 3.44,$$

-standard deviation (as a measure of dispersion) is about 2.46,

- the skewness coefficient is approximately 0.8, which indicates that the distribution is moderately skewed to the right,

-the kurtosis is approximately 2.45 which indicates that the flatness is higher than normal.

Moment generating function of this distribution is of the form

$$M_X(t) = E(e^{tX}) = \sum_{x=1}^{9} e^{tx} log_{10}\left(1 + \frac{1}{x}\right)$$

$$= \sum_{x=1}^{9} log_{10}\left(1 + \frac{1}{x}\right)^{e^{tx}}$$

$$= log_{10} \prod_{x=1}^{9} \left(1 + \frac{1}{x}\right)^{e^{tx}}$$

$$= \frac{1}{ln10} ln \prod_{x=1}^{9} \left(1 + \frac{1}{x}\right)^{e^{tx}}.$$

Next interesting result tells us how to generate Benford random variable. The result is given in the next theorem, Berger & Hill (2021).

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

The main subject of the paper is an overview of distributions of claims used in non-life insurance, that conform to Benford's law. The conformity is confirmed both theoretically (by proving the theorems) and empirically (by conducting simulations). Some of that distributions, which obey the law are: Log-normal, Weibull, Inverse gamma and Inverse Gaussian.

The next theorem is used in the proofs that Log-normal and Weibull distributions obey Benford's law. That result can be find in the literature, Fang & Chen (2020).

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

Below is a result which gives the expression of the distribution of the fraction of random variable from the Log-normal distribution and measure the deviation from the unifom distribution in $[0,1]$.

Theorem 3: Let $X_{\mu,\sigma}$ be a random variable drawn from the Log-normal distribution with parameter μ and σ. For $z \in [0,1]$, let $F_B(z)$ be the cdf of $\langle log_B(\mathrm{X})\rangle$. Then $F_B'(z)$ can be expressed as

$$F_B'(z) = 1 + \sum_{k\geq 1} \cos\left(zk - \frac{2\pi\mu k}{\log B}\right) \cdot \exp(-\frac{2\pi^2\sigma^2k^2}{(\log B)^2}).$$

For these computation in the previous theorem it was used Fourier analysis and Poisson summation formula, and it is not a trivial calculation. As the difference between $F_B'(z)$ and 1 measures the deviation from Benford's law, it is useful to have a good estimate for the sum over k. This results and complite proof can be find in Fang & Chen (2020).

Also, similar result for Weibull distribution is given in Cuff et al. (2015) and we give only formulation of the theorem, without the proof.

Theorem 4: Let $Z_{\alpha,\gamma}$ be a random variable whose density is a Weibull with parameters α, $\gamma > 0$ arbitrary. For $z \in [0,1]$, let $F_B(z)$ be the cdf of $\langle log_B(\mathrm{X})\rangle$. Then $F_B'(z)$ can be expressed as

$$F_B'(z) = 1 + 2\sum_{m=1}^{\infty} \mathcal{R}\left(e^{-2\pi i m\left(z-\frac{\log\alpha}{\log B}\right)} \cdot \Gamma\left(1 + \frac{2\pi i m}{\gamma \log B}\right)\right),$$

where $\mathcal{R}(z)$ denotes the real part of z and $\Gamma(s) = \int_0^\infty x^{s-1}e^{-x}dx$ is a Gamma function of s, for $\mathcal{R}(s) > 0$.

For Inverse gamma distribution numerical simulations were carried out in order to prove its conformity with the Benford's law. In the literature, for example see Fang & Chen (2020), the numerical simulation is divided into two steps.

The first step is calculation of the probability of the first digit for any random variable $X > 0$ ($D_1(X)$), with continuous distribution function $G(x)$ and density function $g(x)$:

$$P(D_1(X) = d) = \sum_{i=-\infty}^{+\infty} \int_{d\cdot 10^i}^{(d+1)\cdot 10^i} g(x)dx,$$

and also it is possible to write previous equation in terms of the distribution function $G(x)$:

$$P(D_1(X) = d) = \sum_{i=-\infty}^{+\infty} \left[G\left((d+1)\cdot 10^i\right) - G(d\cdot 10^i)\right].$$

The second step is measuring the goodness of fit between the probability distribution of the first digit of X ($D_1(X)$) and Benford's law, based on Chi-square test, J-divergence and correlation coefficient. J-divergence is defined as the sum of two possible Kullback-Leibler distances and it is given with the next formula: $J = KL(P,B) + KL(B,P)$, where

$$KL(P,B) = \sum_x P(x)\cdot \log\left(\frac{P(x)}{B(x)}\right) = \sum_{k=1}^{9} P(D_1(X) = k)\cdot \log\left(\frac{P(D_1(X) = k)}{B(k)}\right),$$

with P and B, the two discrete probability distributions, P is for the first digit of X and B is for the Benford's law. The part $KL(B,P)$ can be obtained with exchanging the role of P and B.

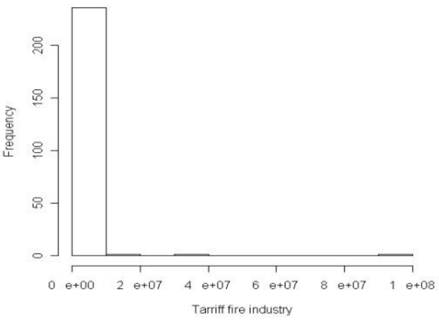
They considered Inverse gamma distribution with diffrent parameters and results indicate that goodness of fit is better if shape parameter decreases, probabiliities of $D_1(X)$ approach the probability of Benford's law, and the scale parameter almost has no effect. The result is if the shape parameter is less than or equal to 0.3, then $J \leq 0.0002$ and correlation coefficient between random variables from these two distributions (Inverse gamma and Benford) is almost equal to 1.

Inverse Gaussian distribution is considered in working paper Stanojevic et al. (2023) and is proved that this distribution also obeys the Benford's law. Complete proof has been given in the above mentioned paper. Also simulations have been used to show the same result.

3. DATA AND METHODOLOGY

We analyzed data about values of losses during the period 1997-2002 in the tariff fire-industry. That series has been converted to the december 2002 constant price series for data comparability. These data were also analyzed in Ćojbašić and Tomović (2007). The histogram of distribution of losses is shown in Figure 2.

Figure 2. *Histogram of data*



Source: *Ćojbašić and Tomović (2007)*

By using the statistical software BestFit, it was obtained that the observed data can be described by Log-normal[1] or Inverse Gaussian[2] distribution, which is shown in Table 1.

Table 1. *Fitted distributions*

Distribution	Log-normal	Inverse Gaussian
Parameters	μ=548906	μ=1076223
	σ=2905044	λ=45226
χ^2 test	19.20	18.00
p value	0.2050	0.2624

Source: *Output from BestFit*

For analysis of validate distribution according to Benford's law (the null hypothesis is that the distribution conformed with the Benford's law) we use four statistical tests: Z-statistic, Chi-square (χ^2) test, KS test and Mean Absolute Deviation (MAD), which were proposed by Nigrini (2012) and Costa et al. (2013). Those tests are completely different. Further, we describe all tests, and give all calculations at 1% of significance level. Z-test checks whether the individual distribution significantly differs from expected Benford's law distribution, for every digit separately. The formula is:

$$Z_i = \frac{|p_{oi} - p_i| - \frac{1}{2n}}{\sqrt{\frac{p_i(1-p_i)}{n}}}$$

where Z_i is Z-statistic for the digit i ($i = 1, 2, \ldots, 9$ for the first digit), p_{oi} is the observed frequency proportion of the digit i, p_i is the expected frequency proportion of the digit i according to the Benford's law, n is the number of observations of the examined variable, the term $\frac{1}{2n}$ is Yates' correction factor and it is used when it is smaller then the absolute difference $|p_{oi} - p_i|$ in the numerator. If the value of Z-statistic exceeds the critical value 2.575, the null hypothesis is rejected at 1% of significance level. Chi-square test is used to test the difference of the "whole distribution" of the observed frequencies of the first digits and the distribution of their expected frequencies under the Benford's law. Advantage in comparation with z-test is that this test is conducted over all digits at the same time (simultaneously). If the test rejects the null hypothesis then this is a signal for data manipulation and it is recommended to look deeper into the data. The Chi-square statistic is calculated as is shown in the next formula:

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

for the first digit test, where O_i is the observed frequency of the digit i, E_i is the expected frequency of the digit i implied by the Benford's law ($E_i = np_i$). The number of degrees of freedem for statistic χ^2 is 8. Observed value of Chi-square test statistic is compared to a critical value, which is 20.090 and the higher it is, the more data deviates from the Benford's law. Once it exceeds, the null hypothesis is rejected.

Appropriate KS statistic is given with the next formula, for testing the first digit:

$$KS = \frac{1}{\sqrt{n}} max_{1 \le j \le 9} \,|\sum_{i=1}^{j}(O_i - E_i)|$$

[1] Log-normal distribution (denoted by Log-norm(μ,σ)) is defined by its pdf: $f(x) = \frac{1}{x\sqrt{2\pi}\sigma'}\exp\left(-\frac{1}{2}\left[\frac{\ln x-\mu'}{\sigma'}\right]^2\right)$, $0 < x < +\infty$, where $\mu' = \ln[\frac{\mu^2}{\sqrt{\sigma^2+\mu^2}}]$ and $\sigma'^2 = \ln[1 + \left(\frac{\sigma}{\mu}\right)^2]$. Expected value and variance of a random variable with Log-normal distribution are μ and σ^2, respectively.

[2] Inverse Gaussian distribution (denoted by IG(μ, λ)) is defined by its pdf: $f(x) = \sqrt{\frac{\lambda}{2\pi x^3}} \cdot e^{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}}, 0 < x < +\infty$. Expected value and variance of a random variable with Inverse Gaussian distribution are μ and $\frac{\mu^3}{\lambda}$, respectively.

$$= \sqrt{n}\; max_{1 \le j \le 9} \left| \sum_{i=1}^{j} (p_{oi} - p_i) \right|$$

where O_i, E_i, p_{oi}, p_i, i and n were introduced earlier.

Mean Absolute Deviation (MAD) test ignores the dataset size, as previous test considered, and it is indicated to large databases, Nigrini (2012). The statistic is calculating with the next formula:

$$MAD = \frac{1}{n} \frac{\sum_{i=1}^{9} |O_i - E_i|}{9} = \frac{\sum_{i=1}^{9} |p_{oi} - p_i|}{9}$$

where O_i, E_i, p_{oi}, p_i, i and n were introduced earlier. There are no objective critical scores for the MAD test. Nigrini (2012) suggested critical scores for conformity, for nonconformity, and for some in-between categories based on personal experience, what is the lack of this test. Range of MAD critical values for testing the first digit is given in the next table.

Table 2. *Range of MAD critical values for the first digit*

Range	**First digit**
Conformity	0.000-0.006
Acceptable conformity	0.006-0.012
Marginally acceptable conformity	0.012-0.015
Nonconformity	Above 0.015

Source: Nigrini (2012)

4. EMPIRICAL ANALYSIS

Table 3 presents the results for the values of losses in the tariff fire-industry for one town in Serbia during the period 1997-2002. The observed value of Chi-square test statistic is 16.581 and it is lower than the critical value (which is 20.090 with 8 degree of freedom while appropriate *p*-value is 0.035). This means that we do not reject H_0 and we can not conclude that there are manipulations with these reported values of losses for this company.

Table 3. *Statistics of first digit for losses during the period 1997-2002 in the tariff fire-industry*

Number	**Observed proportion (%)**	**Expected proportion (%)**	**Deviation Rate (%)**	**Z-statistic**
1	37.238	30.1	7.138	2.335
2	13.807	17.61	3.803	1.458
3	13.389	12.49	0.899	0.323
4	10.042	9.69	0.352	0.075
5	10.042	7.92	2.122	1.095
6	5.021	6.7	1.679	0.909
7	2.511	5.8	3.289	2.037
8	2.511	5.12	2.609	1.684
9	5.439	4.58	0.859	0.481
	χ^2 test 16.581	KS test 1.103	MAD test 0.025	
p-value	0.035	0.061		

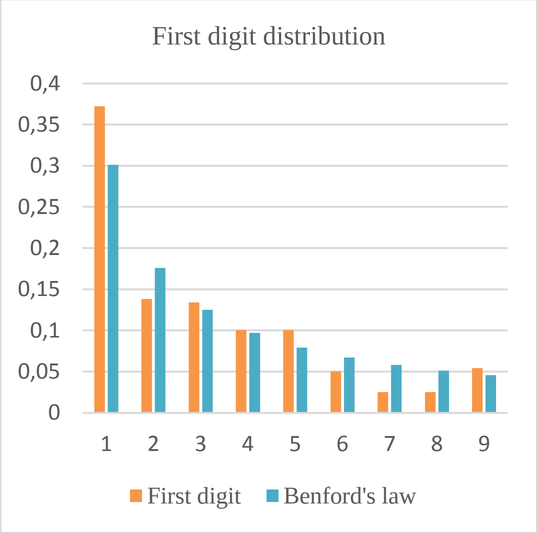
Source: *Authors' calculations*

With z-test we tested all first digits separetly. At 0.01 significance level we do not reject null hypothesis which means that the conformity with Benford's law is not rejected. We can see from Table 3 that for all digits observed value of Z statistic is smaller than critical value.

As KS statistic is 1.103 and appropriate *p*-value is 0.061, that means that null hypotesis of conformity with Benford's law is not rejected (it means there is not an evidence of manipulations with these reported values of losses for this company, in a first digit, testing with KS test).

On the other side, as the value of MAD test is 0.025, that means nonconformity with the law. This conclusion we reached with the MAD test does not necessarily indicate manipulation or fraud, but tells us that a deeper analysis is necessary.

Figure 3. *First digit vs Benford's law frequencies for losses during the period 1997-2002 in the tariff fire-industry*



Source: *Authors' calculations*

In Figure 3 are presented frequencies of the observed data set and expected frequencies according with Benford's law, in a first digit. This figure intuitively shows us the agreement of the data with the Benford's law.

CONCLUSION

In this paper the main goal was to determine the distribution that best fits the considered data, value of losses during the period 1997-2002 in the tariff fire-industry and then to examine the agreement with Benford's law using statistical tests applied in the literature.

Moreover, we obtained that Log-normal and Inverse Gaussian distributions fit well our data. We also showed that considered data obeys Benford's law. This is just a confirmation that data from the Log-normal or Inverse Gaussian distribution follows the Benford's law.

The conclusions of the MAD test should be approached with caution, because the MAD test (due to its high sensitivity) requires additional checks.

Although someone may doubt the validity of the data on reported losses, we have shown in this paper that it doesn't have to be case. The main result of this paper was to determine the agreement or disagreement of the observed data set with Benford's law, thus providing an answer to the question about their possible manipulation and a possible proposal for a deeper analysis of individual numbers.

REFERENCES

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

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

[3] Ćojbašić, V., Tomović, A. (2007). Nonparametric confidence intervals for population variance of one sample and the difference of variances of two samples“, *Computational Statistics & Data Analysis* 51, 5562-5578.

[4] 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*.

[5] Cuff, V., Lewis, A., Miller, S. J. (2015). The Weibull distribution and Benford’s law. *Involve a Journal of Mathematics*, *8*(5), 859-874.

[6] Fang, G., Chen, Q. (2020). Several common probability distributions obey Benford’s law. *Physica A: Statistical Mechanics and its Applications*, *540*, 123129.

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

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

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

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

[11] Stanojević, J., Radojičić, D, Rajić, V., Rakonjac-Antić, T. (2023). Statistical analysis of fitting log-normal and inverse Gaussian distributions with Benford’s law, Working Paper.

[12] Warshavsky, M. S. (2010). Applying Benford’s law in financial forensic investigations. *National litigation consultant’s review*, *10*(2), 1-4