MATHEMATICAL ECONOMICS

No. 8(15) 2012

SOME REMARKS ABOUT BENFORD'S DISTRIBUTION

Mateusz Baryła

Abstract. The aim of the following article is to present some facts about Benford's distribution. Its main focus is on selected descriptors of this distribution (such as mean, variance and skewness) and its two major properties, i.e. base invariance and scale invariance. At the end of the paper some applications of Benford's distribution are presented.

Keywords: base invariance, scale invariance, fraud detection.

JEL Classification: C46.

1. Introduction

From the historical point of view, Benford's distribution is connected with a certain discovery which was made in the nineteenth century by a Canadian astronomer and mathematician – Simon Newcomb. The results of his observations were included in a two-page article published in 1881. In this paper Newcomb states "that the ten digits do not occur with equal frequency must be evident to anyone making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones. The first significant figure is oftener 1 than any other digit, and the frequency diminishes up to 9" (Newcomb, 1881, p. 39).

The same observation, probably independently of Simon Newcomb, was made by another scientist, an American physicist − Frank Benford. The results of his research were published in his paper (Benford, 1938). The analysis of twenty various data sets (the collected data concerned, among other things, addresses, American League baseball statistics, numbers appearing in Reader's Digest articles, mathematical tables from engineering handbooks, including nearly 21,000 observations) led him to draw the conclusion that the ten digits do not occur with the same probability.

Mateusz Baryła
Department of Statistics, Cracow University of Economics, Rakowicka Street 27, 31-510 Kraków, Poland.
E-mail: barylam@uek.krakow.pl

In general, Benford's distribution refers to the probability distribution of the occurrence of significant digits in numbers. In order to explain some terms, it should be stressed that the first significant digit of a certain number is the first non-zero digit, counting from the left side of this number, whereas the second significant digit and further ones can also take 0. For example, the first and the second significant digit of number 0.204 equals 2 and 0, respectively.

The probability that the first significant digit D_1 of a number equals d_1 is calculated according to the following formula:

$$P(D_1 = d_1) = \log_{10}(1 + d_1^{-1}), \tag{1}$$

where: $d_1 \in \{1, 2, ..., 9\}$.

Likewise, the probability for the second significant digit D_2 being d_2 is:

$$P(D_2 = d_2) = \sum_{i=1}^{9} \log_{10}(1 + (10i + d_2)^{-1}) \tag{2}$$

where: $d_2 \in \{0, 1, ..., 9\}$.

The presented problem can be generalized. Let D_k be the k-th significant digit of a number. In this case, the probability that D_k equals d_k is given by the equation (3).

$$P(D_k = d_k) = \sum_{i_1=1}^{9} \sum_{i_2=0}^{9} \dots \sum_{i_{k-1}=0}^{9} \log_{10}(1 + (i_1 \cdot 10^{k-1} + \dots + i_{k-1} \cdot 10 + d_k)^{-1}) \tag{3}$$

where: $d_1 \in \{1, 2, ..., 9\}$, $d_j \in \{0, 1, ..., 9\}$, $j = 2, 3, ..., k$.

It should be emphasized that it is possible not only to calculate the probability for individual digits, but also to calculate the common probability of two or more digits. So as to do this, the following formula is used:

$$P(D_1 = d_1,\ ...,\ D_k = d_k) = \log_{10}\left[1 + \left(\sum_{i=1}^{k} d_i \cdot 10^{k-i}\right)^{-1}\right], \tag{4}$$

where: $d_1 \in \{1, 2, ..., 9\}$, $d_j \in \{0, 1, ..., 9\}$, $j = 2, 3, ..., k$.

For example, the probability that the first, second and third significant digit of a number equals 1, 3 and 5, respectively, is calculated in the following way: $P(D_1 = 1,\ D_2 = 3,\ D_3 = 5) = \log_{10}(1 + (135)^{-1}) \cong 0,0032.$ In this article only the analysis of individual digits is going to be presented.

2. Some descriptors of random variable *D*$_k$

Formulas given by the equations from (1) to (3) presented in the previous section are probability mass functions of a random variable D_k (k = 1, 2, ...). It takes values either from the nine-element set consisting of the digits: 1, 2, 3, 4, 5, 6, 7, 8, 9 (in the case of the first significant digit distribution) or the ten-element set which, apart from the just listed nine digits, additionally contains digit 0 (in the case of the second significant digit distribution and further ones). The obvious question that should be posed here concerns some descriptors of the random variable D_k. In particular, the numerical values of three descriptors (the mean, variance and skewness) are computed in this paragraph.

In order to calculate the mean and variance of D_k when k equals 1, these expressions are used:

$$E(D_k) = \sum_{i=1}^{9} i \cdot P(D_k = i), \tag{5}$$

$$Var(D_k) = \sum_{i=1}^{9} i^2 \cdot P(D_k = i) - [E(D_k)]^2. \tag{6}$$

In the case of k taking the values greater than 1, the first moment (the mean) and the second central moment of D_k are calculated using the following formulas:

$$E(D_k) = \sum_{i=0}^{9} i \cdot P(D_k = i), \tag{7}$$

and

$$Var(D_k) = \sum_{i=0}^{9} i^2 \cdot P(D_k = i) - [E(D_k)]^2. \tag{8}$$

So as to assess the asymmetry of the probability distribution of D_k, the following formula is used:

$$\gamma = \frac{\mu_3}{[Var(D_k)]^{3/2}}, \tag{9}$$

where μ_3 is the third moment about the mean which in the case of k taking the value 1 is defined as follows:

$$\mu_3 = \sum_{i=1}^{9} [i - E(D_k)]^3 \cdot P(D_k = i), \tag{10}$$

while for k greater than 1, μ_3 is given by:

$$\mu_3 = \sum_{i=0}^{9} [i - E(D_k)]^3 \cdot P(D_k = i). \tag{11}$$

The outcomes of the conducted calculations for D_k, where k changes from 1 to 5, are presented in Table 1. In the second and third column of the table, both the means and variances are given in brackets if the distribution is uniform.

Table 1. Mean, variance and skewness of D_k for selected k

k	$E(D_k)$	$Var(D_k)$	$\gamma(D_k)$
1	3.44023696712 (5.0)	6.05651263 (6.67)	0.7956
2	4.18738970693 (4.5)	8.25377862 (8.25)	0.1331
3	4.46776565097 (4.5)	8.25009436 (8.25)	0.0137
4	4.49677537552 (4.5)	8.25000095 (8.25)	0.0014
5	4.49967753636 (4.5)	8.25000001 (8.25)	0.0001

Source: own calculations.

As shown in Table 1, the random variable D_1 has the mean of 3.44 and the variance of 6.06. Taking into account the mean and variance of D_k for k greater than 1, one can observe that these two moments are approaching 4.5 and 8.25, which are the first moment and the second central moment, respectively, if the distribution is uniform. Analyzing the computed values of skewness, one can see that the first significant digit distribution is skewed to the right and the asymmetry is moderate. In addition, the further a significant digit distribution is considered, the weaker asymmetry is observed.

The same conclusions regarding asymmetry can be drawn using the graphical presentation of probability mass function of D_k. Fig. 1 shows the first significant digit distribution, whereas Fig. 2 presents the distribution of further significant digits.

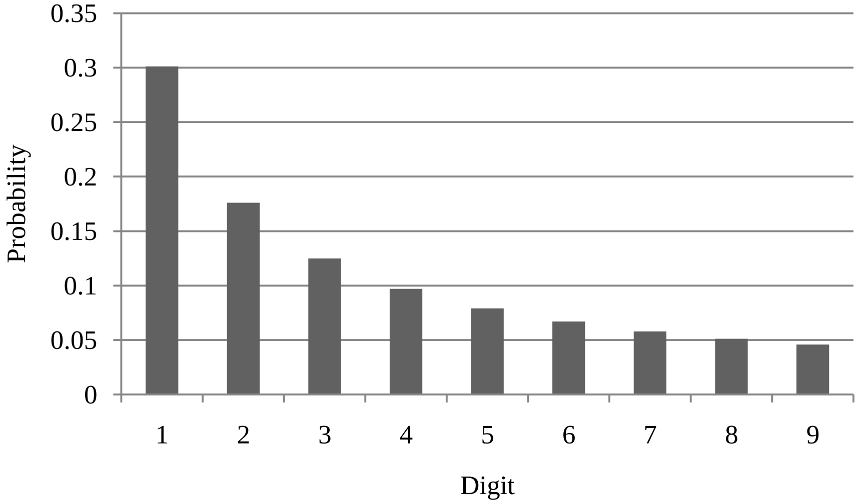

Fig. 1. Distribution of the first significant digit

Source: own calculations.

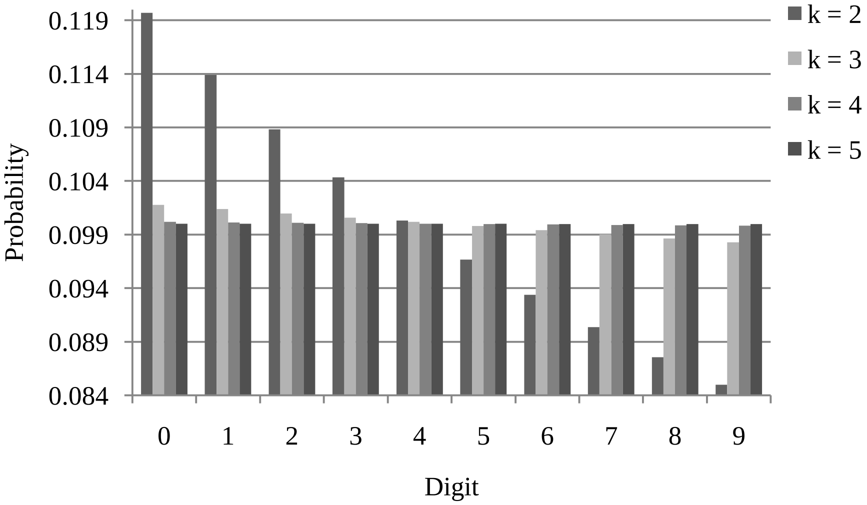

Fig. 2. Distribution of the *k*-th significant digit for *k* = 2, 3, 4, 5

Source: own calculations.

Taking the first digit distribution into account, a moderate positive skewness is noticeable. Moreover, it can be observed that the *k*-th significant digit distribution approaches the uniform distribution when $k \to \infty$.

3. Properties of Benford’s distribution

When discussing Benford’s distribution, it is important to emphasize some of its characteristic properties. In this section, two such properties are going to be discussed − base invariance and scale invariance.

So far all the presented remarks regarding Benford’s distribution have concerned the situation when only the most common numeral system was taken into consideration, i.e. the decimal system. Nevertheless, Benford’s distribution refers also to other numeral systems, which makes this distribution base invariant.

Let b denote a base. Then, for any base $b\,(b > 1 \wedge b \in N)$ the probability mass function for the first, second and − generalizing the problem – k-th significant digit is given by the following equations:

$$P(D_1 = d_1) = \log_b(1 + d_1^{-1}), \tag{12}$$

where: $d_1 \in \{1, ..., b-1\}$;

$$P(D_2 = d_2) = \sum_{i=1}^{b-1} \log_b(1 + (ib + d_2)^{-1}), \tag{13}$$

where: $d_2 \in \{0, ..., b-1\}$;

$$P(D_k = d_k) = \sum_{i_1=1}^{b-1} \sum_{i_2=0}^{b-1} \cdots \sum_{i_{k-1}=0}^{b-1} \log_b(1 + (i_1 b^{k-1} + \ldots + i_{k-1} b + d_k)^{-1}), \tag{14}$$

where: $d_1 \in \{1, ..., b-1\}$, $d_j \in \{0, ..., b-1\}$, $j = 2, 3, ..., k$.

If we substitute 10 for b (i.e. the decimal system is considered), then the above written equations are the same as the ones which were presented in the first section (see equations (1), (2) and (3)).

As an example, let us concentrate on the first significant digit distribution, but this time taking a different base. The computed probabilities for various bases are presented in Table 2. Additionally, Fig. 3 shows the probability distribution for selected even bases from 2 to 10.

Table 2. Distribution of the first digit for bases from $b = 2$ to $b = 10$

b d_1	10	9	8	7	6	5	4	3	2
1	0.30103	0.31546	0.33333	0.35621	0.38685	0.43068	0.50000	0.63093	1.00000
2	0.17609	0.18454	0.19499	0.20837	0.22629	0.25193	0.29248	0.36907	
3	0.12494	0.13093	0.13835	0.14784	0.16056	0.17875	0.20752		
4	0.09691	0.10156	0.10731	0.11467	0.12454	0.13865			
5	0.07918	0.08298	0.08768	0.09369	0.10176				
6	0.06695	0.07016	0.07413	0.07922					
7	0.05799	0.06077	0.06422						
8	0.05115	0.05361							
9	0.04576								

Source: own calculations.

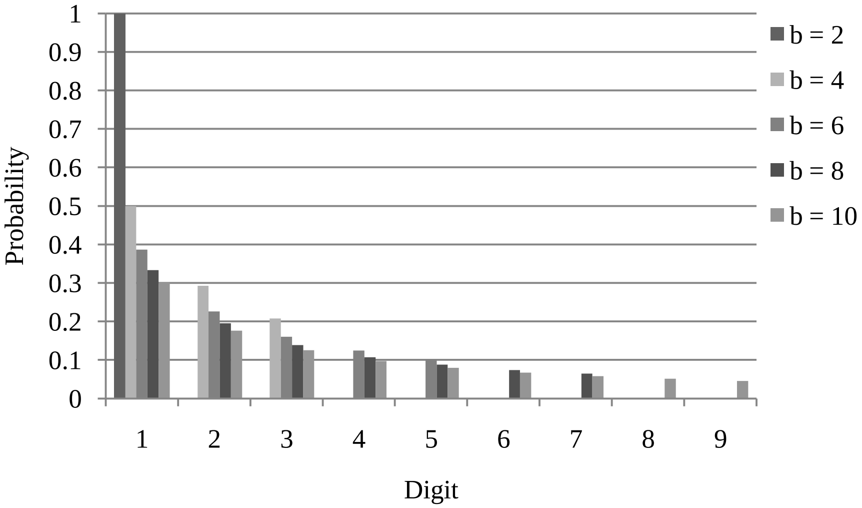

Fig. 3. Distribution of the first significant digit for various bases

Source: own calculations.

The binary system (b = 2) is an interesting case here. In this situation, the probability that the first significant digit is 1, equals 1. It is a well-known fact that in the binary system only two values are possible, i.e. 0 and 1. Remembering that digit 0 can never be the first significant one, the first digit must be 1. When discussing the decimal system (b = 10), one can notice that about 30.1% of all numbers have digit 1 as the first significant digit, whereas almost 4.6% of them have digit 9.

Furthermore, not only is Benford's distribution base invariant but also scale invariant. From a practical point of view, this property means that if e.g. the first significant digit distribution of a certain data set stays in accordance with Benford's distribution (in literature this kind of data set is sometimes called the Benford set), then multiplying each element of this data set by a positive constant leads to a new Benford set, as was shown by Roger Pinkham (see Pinkham, 1961). Despite conducting the mathematical proof of this property in the aforementioned paper, a simple empirical example explaining this characteristic is presented below.

In order to create a data set, six hundred numbers were taken from the Statistical Yearbook of the Malopolska Voivodeship (2009) at random. Then, the data set (data set A) was analyzed with special attention to the first significant digit distribution. Fig. 4 shows the observed relative frequencies of each digit. The chi-square test based on differences between observed and expected frequencies was used as well, which yielded the following results: $\chi^2 = 1.0386$; df = 8; $p = 0.9980$.

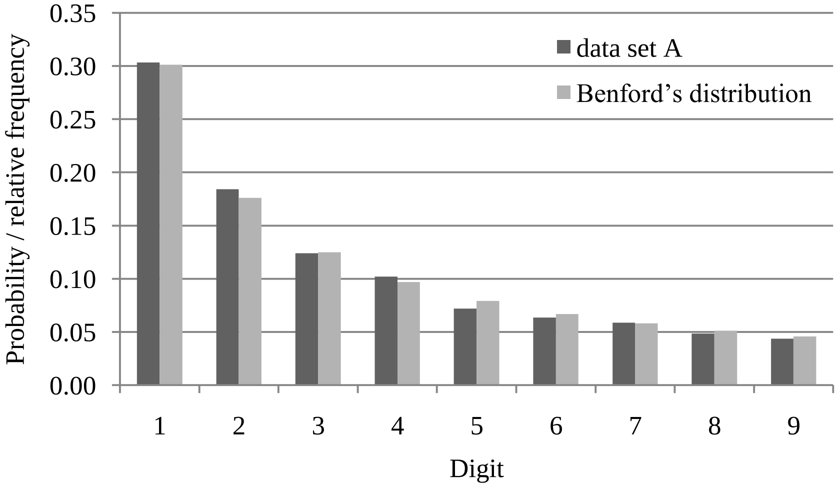

Fig. 4. Distribution of the first significant digit for data set A

Source: own calculations.

The presented outcomes confirm that the data set follows Benford's distribution. Therefore, two mathematical operations were made, which led to the creation of two new data sets (B and C). Data set B is the result of multiplying each element of data set A by π^2, while data set C contains numbers which are the results of multiplying each element of data set A by 7.8. Fig. 5 and 6 present the first digit distribution for the new sets of data. In both cases for α = 0.05 the chi-square goodness-of-fit test did not permit to reject the null hypothesis (for data set B: χ^2 = 1.3949; df = 8; p = 0.9943; for data set C: χ^2 = 4.5504; df = 8; p = 0.8044), stating that the observed distribution of the first significant digit conforms to Benford's distribution.

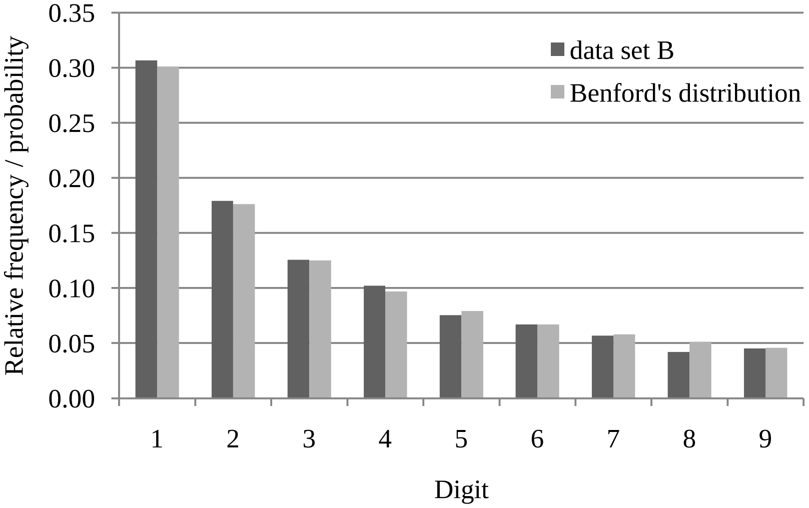

Fig. 5. Distribution of the first significant digit for data set B

Source: own calculations.

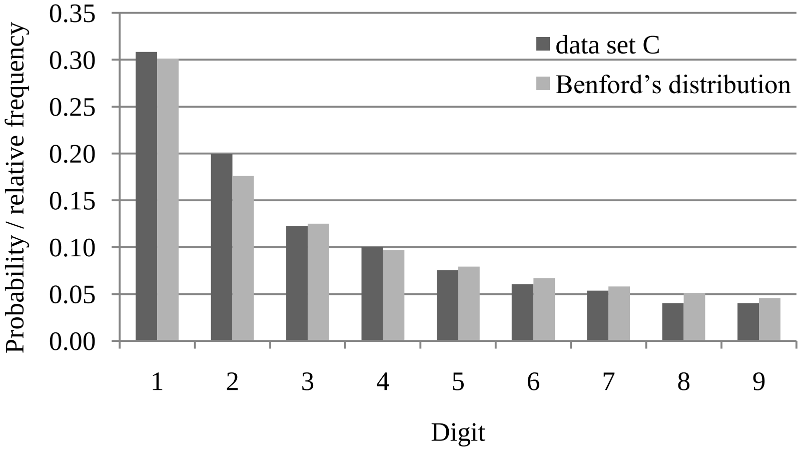

Fig. 6. Distribution of the first significant digit for data set C

Source: own calculations.

Scale invariance seems to be a very important property, especially when converting data expressed in various units. For instance, there is no difference whether share prices are expressed in dollars, yen or any other currency.

4. Applications

On the one hand, Benford's distribution is very interesting as an example of a certain discrete probability distribution. However, while discussing this distribution, it seems important to point out some of its practical implementations as well. Undoubtedly, the most popular application of Benford's distribution is fraud detection. For this reason, this distribution is especially interesting for auditors who try to identify errors in accounting data.

The first papers that concern the accounting application of the distribution appeared in the late 1990s. Two studies derive from this period. Charles Carslaw analyzed earnings numbers of New Zealand companies. In his paper (Carslaw, 1988), he found that managers tended to round these numbers up (if a company had, for instance, earnings equal to 697,000 USD, it would be rounded up to 700,000 USD) because of the fact that the numbers did not conform to Benford's distribution taking into consideration the second significant digit distribution − there were too many 0s and too few 9s. A similar analysis can also be found in the paper (Thomas, 1989). In this case the author carried out his studies using data that concerned earnings of American companies.

Mark Nigrini had a great impact on the application of Benford's distribution in detecting fraud. He broadly described many analytical procedures which can be very useful for auditors when analyzing data sets that contain accounting numbers. Nigrini proposed, among other things, some tests based on the digital analysis such as: first digit test, second digit test, first-two digits test, last-two digits test, etc. The basic information regarding the use of Benford's distribution in practice is included in the following papers: (Nigrini, Mittermaier, 1997; Drake, Nigrini, 2000; Nigrini, 2000).

Detecting fraud in accounting data, however, is not the only application of Benford's distribution. There are also known attempts to use it to help identify tax evaders and check the reliability of held elections.

5. Conclusion

Comments on Benford's distribution presented in the article (i.e. moments and some characteristic properties) have been made with an assumption that Benford's distribution is a discrete probability one. It is also important to emphasize that it is possible to make a generalization of this distribution. Those interested in this problem should take into consideration for example the following paper (Scott, Fasli, 2001).

When conducting empirical analysis, the necessityof checking data conformity with Benford's distribution arises very often. In this situation, commonly applied methods can be used, such as: chi-square goodness-of-fit test, Z-statistic test, Kolmogorov-Smirnov test, Kuiper test, Mean Absolute Deviation or regression analysis. There are also available some modified tests described in (Morrow, 2010).

Literature

Benford F. (1938). *The law of anomalous numbers*. Proceedings of the American Philosophical Society. Vol. 78. No. 4. Pp. 551-572.

Carslaw C. (1988). *Anomalies in income numbers: evidence of goal oriented behavior*. The Accounting Review. Vol. 63. No. 2. Pp. 321-327.

Drake P.D., Nigrini M.J. (2000). *Computer assisted analytical procedures using Benford's Law*. Journal of Accounting Education. No. 18. Pp. 127-146.

Morrow J. (2010). *Benford's Law, families of distributions and a test basis*. http://www.johnmorrow.info/projects/benford/benfordMain.pdf.

Newcomb S. (1881). *Note on the frequency of use of the different digits in natural numbers*. American Journal of Mathematics. Vol. 4. No. 1. Pp. 39-40.

Nigrini M.J. (2000). *Continuous auditing*, Ernst & Young Center for Auditing Research and Advanced Technology. University of Kansas.

Nigrini M.J., Mittermaier L. (1997). *The use of Benford's Law as an aid in analytical procedures*. Auditing: A Journal of Practice & Theory. Vol. 16. No. 2. Pp. 52-67.

Pinkham R.S. (1961). *On the distribution of first significant digits*. The Annals of Mathematical Statistics. Vol. 32. No. 4. Pp. 1223-1230.

Scott P.D., Fasli M. (2001). *Benford's Law: An Empirical Investigation and a Novel Explanation*. CSM Technical Report 349. Department of Computer Science. University of Essex. http://cswww.essex.ac.uk/technical-reports/2001/CSM-349.pdf.

Statistical Yearbook of the Malopolska Voivodeship 2009, Statistical Office in Cracow, Cracow 2009.

Thomas J.K. (1989). *Unusual patterns in reported earnings*. The Accounting Review. Vol. 64. No. 4. Pp. 773-787.