ORIGINAL PAPER

USING BENFORD'S LAW IN THE ANALYSIS OF SOCIO-ECONOMIC DATA

DAN-MARIUS COMAN[1*], MARIA-GABRIELA HORGA[2], ALEXANDRA DANILA[2], MIHAELA-DENISA COMAN[3]

Manuscript received: 12.09.2017; Accepted paper: 22.11.2017;
Published online: 30.03.2018.

Abstract: *The article represents a theoretical research which takes into account the statistical model known as Benford's Law, followed by a quantitative application-based research in which we have used Microsoft Excel elements (statistical and graphical functions) on a set of socio-economic data in order to emphasize how easy it is for a wide range of users to apply the theoretical concepts mentioned.*

Keywords: *Benford's Law, statistical test, spreadsheet*

1. INTRODUCTION

The statistical model popularized by Benford [1] in 1938 and developed by Newcomb [8] is researched and discussed in a wide range of articles and highlights the universality of applying the method on a set of data such as: electricity bill payment [2], accounting fraud detection and financial statement audits [7], socio-demographic studies [9]. This research aims to use the statistical model popularized by Benford in order to show how easy it is to apply it to a set of data by utilizing a software application (Microsoft Excel) available in the application package a computer is usually equipped with.

2. OBJECTIVES AND METHODOLOGY OF THE RESEARCH

The article represents a positive research approach, in which the theoretical concept is reflected and practically demonstrated by an example in the socio-economic field. The method employed in this research was the consultation of the specialized literature presented in the bibliography, along with the development and detailing of the theoretical concept by means of a practical example. Through its practical nature, the analytical procedure based on the quantitative model of Benford's law highlights a statistical relationship and falls under the pragmatic research.

[1] Valahia University of Targoviste, Faculty of Economic Sciences, 130004 Targoviste, Romania.
* Corresponding author E-mail: cmnmarius@yahoo.com.
[2] Ovidius University of Constanta, Faculty of Economic Sciences, 900470 Constanta, Romania.
E-mail: alexandradanila14@yahoo.com , gabi_horga@gmail.com.
[3] Valahia University of Targoviste, Institute of Multidisciplinary Research for Science and Technology, 130004 Targoviste, Romania. E-mail: cmndenisa@yahoo.com.

3. MATHEMATICAL FOUNDATIONS OF BENFORD'S LAW

Benford's law refers to the frequency distribution of digits in various situations of occurrence of numerical data within real-life sources of data. The digit 1 occurs as leading digit in 30% of the cases, while greater digits occur with smaller probability. The distribution of first digits is as large as the interval of a logarithmic scale and the results apply to a wide variety of data. Mathematically, the distribution law applies to base-10 numbers, but there are applications of the law for the distribution of other numbers in other bases or for the second or the following digits of a number.

A number string follows Benford's distribution if the distribution probability of one digit (d_i) of the numbers ($d=d_0d_1.....$) represented in the in the b-base number system satisfies the law:

$$p(d_0) = log_b\left(1 + \frac{1}{d_0}\right), d_0 = 1 \dots (b-1) \quad (1)$$

$$p(d_1) = \sum_{k=1}^{b-1} \log b\left(1 + \frac{1}{k * b + d_1}\right), d_1 = 0 \,..\, (b-1) \quad (2)$$

$$p(d_2) = \sum_{j=1}^{b-1}\sum_{k=0}^{b-1} \log_b\left(1 + \frac{1}{j * b^2 + k * b + d_2}\right), d_2 = 0 \,..\, (b-1) \quad (3)$$

Numerically, the first digits of a number following Benford's law have the distribution in Table 1.

Table 1. Data distribution according to Benford's law.

Digit	Probability
1	0.30103
2	0.17609
3	0.12494
4	0.09691
5	0.07918
6	0.06695
7	0.05799
8	0.05115
9	0.04576

The probability p(d) is proportional to the width of the interval between d and d+1 on a logarithmic scale, so it will comply with the expected distribution of the mantissa of the logarithm of that particular number, but not of that number proper, as it is probabilistically uniformly distributed. The distribution law is named after the physicist Frank Benford who formulated it intuitively in 1938. According to [3], Benford's law was explained in several ways:

- Consequence of the exponential growth process: in principle, it is based on the assumption that the mantissa of logarithms of numbers is uniformly distributed. This is correct if the numbers are themselves distributed on multiple orders of magnitude;
- Scalar invariance: the distribution of digits in a real list does not depend on the unit of measurement used, in other words, multiplication by the same constant will not affect

distribution. The phenomenon is called scalar invariance and the variables which are log-normally distributed comply with this property;

- Multiple probability distributions.

The model of Benford's law applies to datasets which are distributed on several orders of magnitude. A direct consequence resulting from this observation is that the model is not valid if we want to check the values in a list (for example bills, payments) between two limit values (for example between 50,000 and 100,000) or above a minimal value or below a maximal value.

The particular criteria [10] which limit the application of Benford's law have been emphasized on accounting datasets and refer to:

- Assigning numbers for bills, cheques;
- Influencing numbers by human subjective decisions (prices like 9.99);
- Accounts with maximum and minimum limits.

4. PRACTICAL APPLICATION

In addition to applications in the field of fundamental sciences [5] Benford's law has found an application in fraud detection in all economic activities. That is why the model has been included in the CAATS (Computer Auditing Techniques) analytical procedures.

In this article, Benford's model is applied in order to detect abnormal results in a list which include data related to the number of people in small towns of the United States of America [11]. The main goal of applying Benford's model in the dataset established is determined by the verification of the following research hypothesis:

"The chosen dataset follows the distribution of Benford's law."

The dataset comprises 3,141 observations on the number of people in small towns located in the United States of America. For space reasons, Fig. 1 presents only a screenshot of the list used in the analysis.

1990 Census Data	Area	Population
Autauga County	73.3	34222
Baldwin County	88	98280
Barbour County	32.8	25417
Bibb County	33.4	16576
Blount County	79	39248
Bullock County	18.7	11042
Butler County	27.5	21892
Calhoun County	184.5	116034
Chambers County	61.3	36876
Cherokee County	43.4	19543
Chilton County	57	32458
Choctaw County	17.4	16018
Clarke County	22.5	27240
Clay County	23.6	13252
Cleburne County	25.2	12730
Coffee County	64.2	40240
Colbert County	92.5	51666
Conecuh County	16.6	14054
Coosa County	18.7	11063
Covington County	36.4	36478
Crenshaw County	22.4	13635
Cullman County	104.9	67613
Dale County	87.6	49633
Dallas County	47.3	48130

Figure 1. Screenshot of the dataset.

Benford's law has been applied in Microsoft Excel, which has options (functions, graphs) to make the necessary calculations. Figures 3 and 4 show the tables in which all the operations required by this analytical method have been performed.

The calculation of the frequency of occurrence of each digit from 1 to 9 is done in Table 2.

The formulas which have determined the calculation of the occurrence frequency of each digit are the following:

• Determining the first digit of numbers in the data series FLOOR(10^MOD(LOG(Data!A8,10),1),1);

• Creating a contingency table in which every digit determined at the previous point is represented by the digit 1 on the position corresponding to the digit sequence in the table header IF(ISERROR($B5) = TRUE,0,IF($B5=3,IF($A5 <= C1,1,0),0));

• Determining the sum for each digit in the table header SUM(D5:D4004).

Table 2. Determining the frequency of occurrence of each digit.

Frequency of occurrence of digits (1 ..9)	**992**	**534**	**424**	**260**	**228**	**210**	**173**	**175**	**145**
Digits 1…9 / **First digit extracted from the data series**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
3	0	0	1	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	1
2	0	1	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0
1	1	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0

The elements determined at 3 in the previous paragraph help us make the table in which we shall apply the elements of Benford's law. The table of analysis is illustrated in Table 3.

Table 3. Determining the elements of Benford's law.

Digit	Observed Probability	% of Observed Probability	Expected Probability (Benford models)	% of Expected Probability (Benford models)	Variation (%)
1	992	31.58	946	30.10	4.68
2	534	17.00	553	17.61	-3.58
3	424	13.50	392	12.49	7.45
4	260	8.28	304	9.69	-17.07
5	228	7.26	249	7.92	-9.08
6	210	6.69	210	6.69	-0.13
7	173	5.51	182	5.80	-5.29
8	175	5.57	161	5.12	8.19
9	145	4.62	144	4.58	0.88
Total	**3141**	**100**	**3141**	**100**	

In Table 3, the frequency of occurrence of the first digit has been inserted from Table 2 by copy-paste and, based on this information, we have determined:

- The proportion of elements observed in the total sample, *B2/B11*100*;
- The number of expected elements according to Benford's law, *LOG((A2+1)/A2,10)*B11*;
- The proportion of expected elements according to Benford's law, *(LOG((A2+1)/A2,10))*100*;
- The variation of observed elements in relation to the expected elements, *((B2-D2)/B2)*100*.

At a first view of the Table 3, one may note that the greatest variation between the probability of observed elements and the probability of expected elements according to Benford's law manifests for digit 4 (-17.07%). This is also shown by Fig. 2, in which we see that the statistical distribution of the first digit follows the theoretical distribution in a similar manner; we should, however, mention that in the case of the digit 4 there is a greater variation between statistical distribution and theoretical distribution.

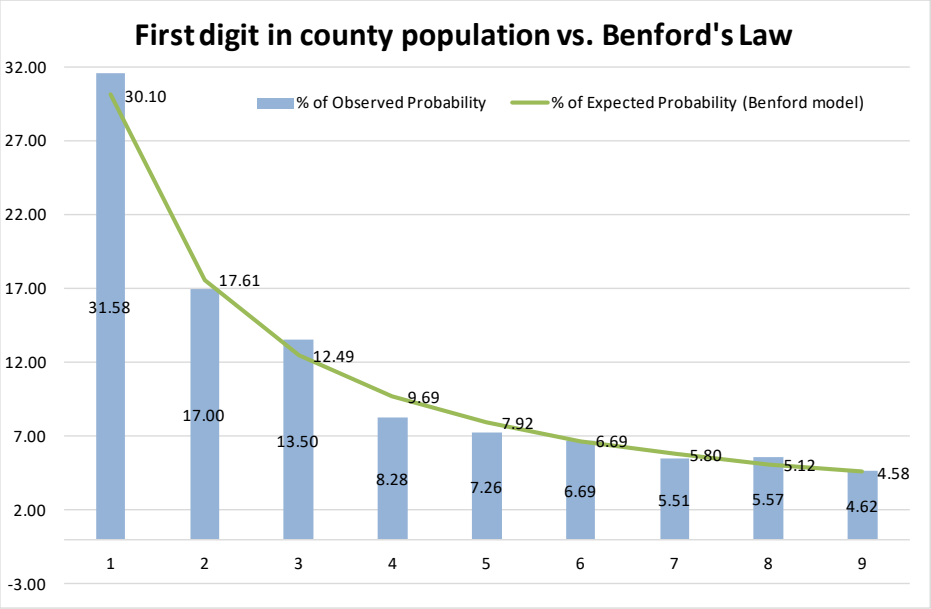

Figure 2. First digit vs. Benford's Law.

The observation makes us consider the research question: **the chosen dataset follows the distribution of Benford's law**, which we can transpose in the following hypothesis:

(1) H_0: The first digit of the dataset on the observed population **follows** the theoretical distribution of Benford's law

(2) H1: The first digit of the dataset on the observed population **does not follow** the theoretical distribution of Benford's law

To test this hypothesis, we shall use the chi-square test [4] which allows the comparison between two distributions (observed vs. theoretical) with a view to establishing either an independence between them or their homogeneity. The chi-square test formula is:

$$chi\ square = \sum \frac{(observed\ count - expected\ count)^2}{expected\ count} \quad (4)$$

The application of the calculation formula was done in table 4, obtaining a value of 15.43. The obtained value is interpreted by comparing it with the corresponding value in the chi-square table. Since in this case there are nine variants, the number of degrees of freedom is c-1, i.e. 8.

Table 4. Determining the elements of the chi-square test.

Digit	Observed Probability	% of Observed Probability	Expected Probability (Benford models)	% of Expected Probability (Benford models)	Variation (%)	Chi square
1	992	31.58	946	30.10	4.68	2.28
2	534	17.00	553	17.61	-3.58	0.66
3	424	13.50	392	12.49	7.45	2.54
4	260	8.28	304	9.69	-17.07	6.47
5	228	7.26	249	7.92	-9.08	1.72
6	210	6.69	210	6.69	-0.13	0.00
7	173	5.51	182	5.80	-5.29	0.46
8	175	5.57	161	5.12	8.19	1.28
9	145	4.62	144	4.58	0.88	0.01
Total	**3141**	**100**	**3141**	**100**		**15.43**

The chi-square value of 15.43 is less than that in the Table 4 (15.5073) at 8 degrees of freedom and a significance threshold p = 0.9847, calculated based on the Microsoft Excel function CHISQ.DIST(G11,8,TRUE).

Based on chi-square distribution, the probability to reject the null hypothesis is calculated. Usually the research question is accepted if the probability (p) of rejecting the hypothesis H_0 is less than 5%. In the context of p=0.9847 > 0.05, which denotes that ***there are no significant differences between the observed distribution and the theoretical one*** and there is no sufficient evidence to believe that the variation observed in the case of digit 4 between the two distributions may be due to the occurrence of a mystification (alteration) of data reporting regarding the town population.

In the specialized literature [6] a number of problems are reported when comparing observed and theoretical distributions. Among the observations noted are: issues regarding the choice of the number of classes, issues on the width of frequency classes, issues regarding the number of observations inside each frequency class. All these elements lead, in the case of Benford's law, to the calculation of an additional test called mean absolute deviation (MAD), which measures the average of absolute deviation of the frequencies of each digit from

Benford's ideal frequency. In this application, MAD is calculated in Table 5 based on the following formula:

$$MAD = \frac{\sum_{i=1}^{k} |observed\ count - expected\ count|}{expected\ count} \quad (5)$$

The value obtained by applying the formula may fall into three classes according to the following algorithm: 0.000 < MAD >0.006 – close conformity; 0.006 < MAD < 0.012 – acceptable conformity; 0.012 < MAD < 0.015 – marginal conformity; MAD > 0.015 – non conformity.

Table 5. MAD Calculation

Digit	Observed Probability	% of Observed Probability	Expected Probability (Benford models)	% of Expected Probability (Benford models)	Variation (%)	Chi square	MAD
1	992	31.58	946	30.10	4.68	2.28	1.48
2	534	17.00	553	17.61	-3.58	0.66	0.61
3	424	13.50	392	12.49	7.45	2.54	1.01
4	260	8.28	304	9.69	-17.07	6.47	1.41
5	228	7.26	249	7.92	-9.08	1.72	0.66
6	210	6.69	210	6.69	-0.13	0.00	0.01
7	173	5.51	182	5.80	-5.29	0.46	0.29
8	175	5.57	161	5.12	8.19	1.28	0.46
9	145	4.62	144	4.58	0.88	0.01	0.04
Total	**3141**	**100**	**3141**	**100**		**15.43**	**0.6625**

Following MAD calculation, the value 0.6625 is classified as non-conformity, which supports the idea that ***there are no significant differences between observed and theoretical distributions***, which has also been shown by the chi-square test.

5. CONCLUSIONS

Benford's law is a powerful tool, integrated into dedicated applications (e.g. Idea Casewear), for auditors in identifying the risk of economic fraud, but the example presented is not intended to highlight abnormal aspects in reporting socio-economic data on the number of inhabitants of towns in the USA. This paper aims to present, in a simple way, how to use an easy working tool in order to calculate specific elements of Benford's law: chi-square test, mean absolute deviation, the graph of correlation of the researched data series with the expected data series.

For regular users, beginning practitioners, who do not have access to specialized applications, the example presented may be a good start in acquiring the skills of using Microsoft Excel spreadsheet to apply statistical tests for fraud detection in judicial accounting expertise or in auditing.

REFERENCES

[1] Benford, F., *Proceedings of the American Philosophical Society*, **4**(78), 551, 1938.
[2] Christian, C., Gupta, S., Lin, S. M., *National Taxa Journal*, **4**(46), 487, 1993.
[3] Coracioni, A., *Audit Financiar*, **11**(104), 23, 2013.
[4] Florea, N.V., Mihai, D.C., *Journal of Science and Arts*, **1**(38), 81, 2017.
[5] Jäntschi, L., Bolboaca, S., Stoenoiu, C., *Bulletin UASM Agriculture*, **66**(1), 82, 2009.
[6] Jäntschi L., *Proceedings of International Conference on Recent Achievements in Mechatronics, Automation, Computer Science and Robotics*, 239, 2011.
[7] Marcini, S., Hamilton T., *Journal of the Risk and Uncertainty*, **1**(32), 57, 2006.
[8] Newcomb, S., *American Journal of Mathematics*, **4**, 39, 1881.
[9] Sandron, F., *Population*, **57**(4-5), 755, 2002.
[10] Shalini, T., Kinjal, M., *IOSR Journal of Economics and Finance*, 1-9, 2014.
[11] https://introductorystats.wordpress.com/2011/11/25/benfords-law-and-us-census-data-part-ii/," last accessed 27.01.2018.