TADEUSZ GRABIŃSKI

# Benford's Law

## Introduction

When we type into Google the phrase ***Benford***, and when our search is limited to Polish, we obtain over 31 thousand hits (as of September 1st, 2008). The first 300 hits pertain mainly to construction machines (fully rotational road carriage), the science fiction writer Gregory Benford and the author of English text books Michael Benford.

The phrase ***Frank Benford*** in Polish results in merely 13 hits; 3 hits are references to the Scandinavian version of wikipedia, 2 are outdated links, and the remaining 8 contain information about Benford's Law obtained from scientific publications[1].

The term ***Benford's Law*** in Polish results in about 50 hits. Among them we can find links to several publications pertaining to Benford's Law as well as links to web sites that present and discuss the following issues: paranormal activity, math trivia, text book information, and teaching materials pertaining to Benford Law[2].

---

[1] Wołowik P., *Matematyka pomaga odkryć fałszerstwa w zeznaniu podatkowym*, Rzeczpospolita, 20.04.2005, http://www.huby.seo.pl/D3/index.php?option=com_content&task=view&id=40&Itemid=3; „Gazeta Wyborcza", 05/04/23 http://www.kpk-ottawa.org/sip/bez_ogonkow/biuletyn/2005/0506.html; Karpiński M., *Prawo Benforda*, „Matematyka w Szkole", No. 31, styczeń/luty /2008.

[2] Paulos J.A., *Matematyk gra na giełdzie*, Wydawnictwo CDW, Azymut-Book 2007. Wołowik P**.** *Prawo Benforda – testowanie i weryfikacja poprawności danych pomiarowych***,** „Przegląd Telekomunikacyjny – Wiadomości Telekomunikacyjne", 11/2005.

As can be expected, Google search provides a substantially greater number of links and references for similar phrases in English. The word ***Benford*** results in approximately 887 thousand hits. Among the first 300 hits, nearly half pertain to Benford's Law. The remaining hits pertain either to persons with that particular name, or to construction machines. The phrase ***Frank Benford*** results in 1800 hits while, in contrast, Benford's Law results in over 60 thousand hits[3].

This brief analysis pertaining to the web search of a particular subject (i.e. Benford) reveals that the knowledge of Benford's Law in Poland is insignificant; meanwhile we are able to find numerous publications about this particular subject in English.

The first scholar who considered this interesting problem was Simon Newcomb, astronomer and mathematician. In the early 1880s he observed that volumes containing logarithmic tables had much dirtier pages in the front than in the back. Basically this meant that, for unexplained reasons, scholars who studied these tables utilized small numbers rather than large numbers. In his work Newcomb provides even a formula according to which we would be able to approximate the frequency of occurrence of first significant digits.

Newcomb's observation went unnoticed for 60 years. In 1938 the physicist, Frank Benford, empirically verified the correctness of Newcomb formula on 20 sets of numbers that contained over 20 thousand digits (among others rivers surface, city population, death rates). Nevertheless, Benford did not provide an explanation as to why we observe a decreasing frequency of occurrence of numbers, in given sets, which start with gradually increasing digits. Not until 1995 the mathematician Thomas Hill managed to prove the essence of this regularity and, in addition, presented its properties and conditions.

Hence it is uncertain whether the law pertaining to the frequency distribution of the first significant digits should be named after Benford, Hill or Newcomb, or possess a totally different definition. In reality there are serious doubts regarding the general functioning of this law; in other words, is it fully justifiable to use the term law, or should we use the term 'regularity' instead[4].

---

[3] Top links according to Google:
http://en.wikipedia.org/wiki/Benford%27s_law
http://mathworld.wolfram.com/BenfordsLaw.html
http://www.intuitor.com/statistics/Benford%27s%20Law.html
http://www.mathpages.com/home/kmath302/kmath302.htm
www.rexswain.com/benford.html

[4] Among other interesting laws that possess a similar character are Estoup-Zips Law (frequency of words occurrence in texts), Heap's Law (which describes the interrelation between text size and the number of words used) and Lotka Law (pertaining to frequency of references in scientific publications depending on the number of articles for a given autor).

## 1. Benford Distribution

According to Benford's Law, the frequency (P) of occurrence of **first** significant digits $D_1$ in **multi-digit** numbers taken from **large** sets of numbers is given by the formula below:

$$P(D_1 = d_1) = \log\left(\frac{d_1 + 1}{d_1}\right) = \log(1 + 1/d_1) \qquad (d_1 = 1,2,...,9) \tag{1}$$

Figure 1. Benford's Law – frequency distribution of first significant digits

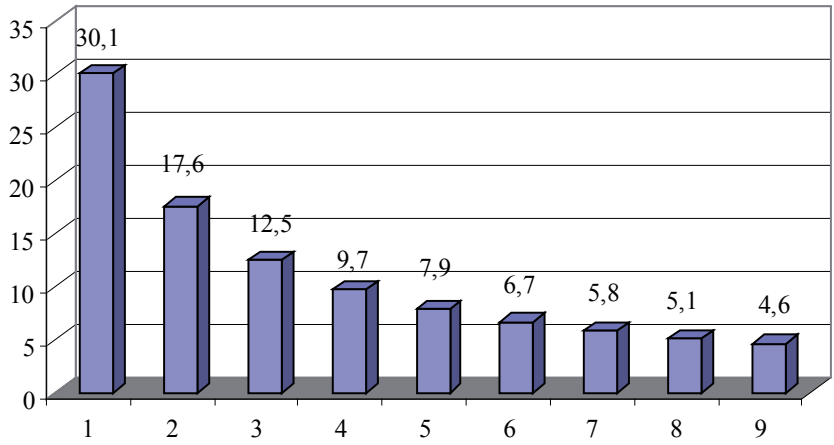
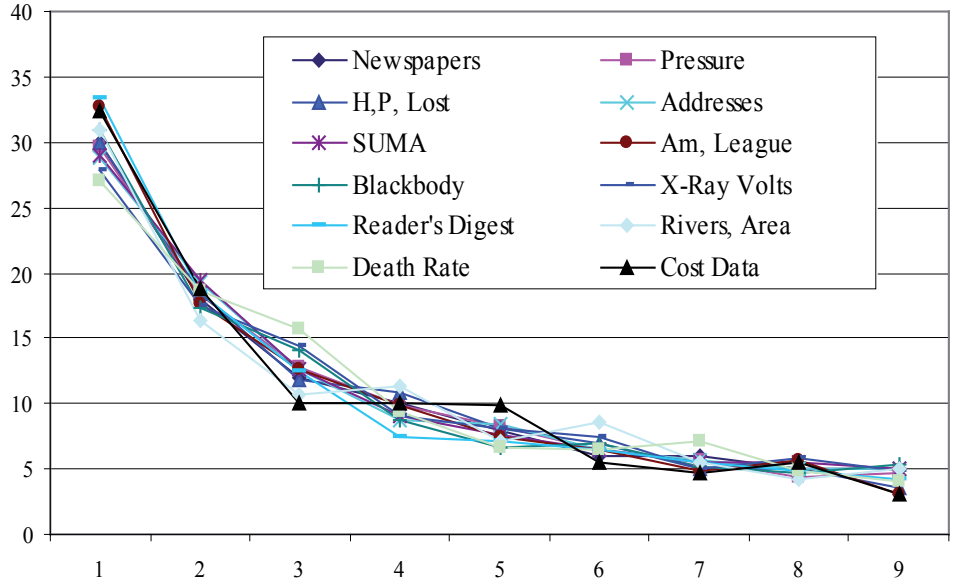



Figure 1 displays the distribution frequency of first significant digits. As can be seen, 1 occurs with a 30% frequency, and subsequent digits have a progressively decreasing frequency (i.e. 9 occurs only with 4,6% frequency).
Figure 2 displays the frequency of first significant digits for 12 separate datasets that have been analyzed by Frank Benford with the highest conformity between empirical and theoretical distributions.

Figure 2. Frequency distributions of first significant digits for 12 datasets analyzed by F. Benford




Benford's Law is precisely associated with the Fibonacci sequence and less familiar Lucas sequence. The above mentioned sequences can be expressed by a recurrent formula presented below:

$$F(n+1) = F(n) + F(n-1)$$

– Fibonacci sequence {1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144,…}
− Lukas sequence {1, 3, 4, 7, 11, 18, 29, 47, 76, 123, 199,…}

Distribution analysis in these sequences, which comprised 1474 numbers, revealed that in both cases (i.e. sequences) the distributions of first significant digits are entirely consistent with Benford distribution (compare with table 1).

Table 1. Distributions for first signicant digits in Fibonacci and Lukas sequences including Benford distribution n=1474

| D | Benford | Fibonacci | Lukas |
|---|---|---|---|
| 1 | 444 | 444 | 445 |
| 2 | 260 | 260 | 258 |
| 3 | 184 | 184 | 185 |
| 4 | 143 | 143 | 143 |
| 5 | 117 | 117 | 116 |
| 6 | 99 | 98 | 99 |

| | | | |
|---|---|---|---|
| 7 | 85 | 85 | 86 |
| 8 | 75 | 76 | 75 |
| 9 | 67 | 67 | 67 |
| **SUM** | **1474** | **1474** | **1474** |

## 2. Attempts to Explain Benford's Law

Scientific literature offers several examples that try to explain the causes of Benford's Law Formula as well as a description of frequency distributions of first significant digits. In this article we will briefly discuss 3 different interpretations of Benford Law.

The first interpretation can be derived from a proven theorem (Thomas Hill), which states that data which are a product of multiple numbers will always be subjected to Benford Law. In reality many types of information possess a product type character. For example, the transaction value of unitary prices to the volume of sales. Hill's theorem is related to the central limit theorem, which deals with sums of random variables and not quotients.

The second interpretation is related to a theorem according to which each sample of randomly mixed numbers from various datasets is subjected to Benford Law. Therefore, if we randomly select numbers from different tables of the statistical yearbook, their first significant digits will be subjected to Benford's Law even if the numbers from individual tables will fail to meet this condition. Based on the above-mentioned remarks, Benford's Law describes the analytical form of "distribution of distributions".

The third interpretation assumes that there exists a general law of nature, which pertains to first significant digits and subsequently reflects the harmonic nature of reality. It should be emphasized that the scale pertaining to sensitivity of sight and sound, as well as many other phenomena (i.e. seismic waves during an earthquake) are often better described by a logarithmic rather than a linear scale.

The intuitive nature of Benford's Law may be best illustrated when we take into account the fact that starting magnitudes possess a unitary level (1, 10, 100, 1000 etc.). If we want to switch to values starting with 2, 20, 200, etc, we have to double the base value (100% increment). Switching from 2, 20, 200,…to 3, 30, 300 requires that base value increase by only 50%. Subsequently switching from 8, 80, 800,… to 9, 90, 900,… increases base value by 12,5%, and from 9, 90, 900,… to 10,100,1000 by 11,1%. It should be noted that the aforementioned increment rates recur cyclically: 100%, 50%, 33%, 25%, 20% etc.

To sum up, in general there are more small and medium size enterprises than large corporations. In a similar fashion, there are more smaller than larger cities, and more small rivers and creeks than large rivers. However, we have to bear in

mind that it may not always be possible to attain subsequent orders of magnitude when starting from uniform magnitude.

## 3. Benford's Law Properties

Benford's Law possesses two essential properties:
(1) Scale non-changeability .
(2) Base non-changeability.
**Scale non-changeability** means that if we multiply, divide or raise data to n-th power by a non zero constant (data characterized by Benford distribution) we will obtain such a distribution that will still be subjected to Benford Law. Therefore, it is irrelevant if magnitudes are expressed in dollars, euro or any other types of units; Benford's Law will still apply to any type of unit. It should be noted that inverse numbers which are subjected to Benford's Law also meet this condition. For instance, the amount of turnover per single share and the amount of shares per 1 dollar worth of turnover.
**Base non-changeability** means that Benford's Law is applicable not just to numbers with a base of 10 but also to other types of number base systems. Thomas Hill proved that Benford's Law is unique since it is the **only** known distribution which possesses this property.
For any given calculation system base (B) the Benford Formula for first significant digits $D_1$, is expressed in the following manner:

$$P(D_1 = d_1) = \frac{\log(1+1/d_1)}{\log(B)} \quad d_1 = (1,2,...,B-1) \tag{2}$$

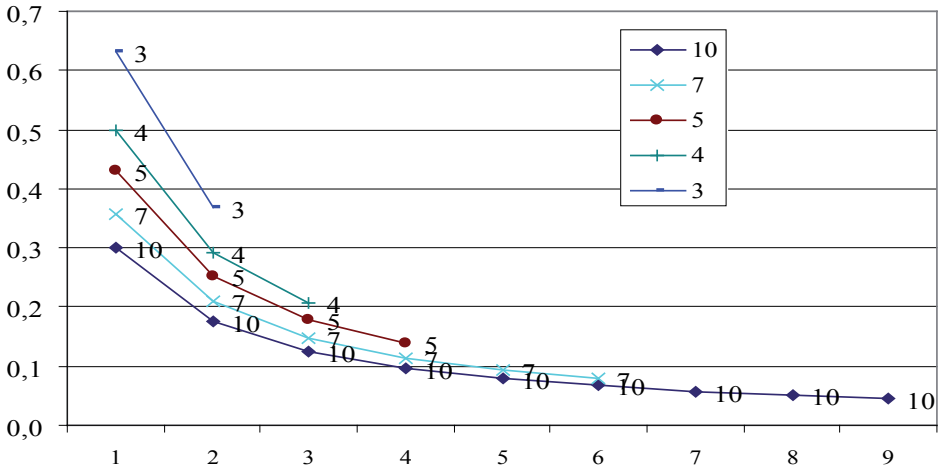
In table 2 and Figure 3 the frequency of occurrence of first significant digits for various number base systems are presented as follows:

Table 2. Frequencies of occurrence of first significant digits for various types of number base systems from B=10 to B=3

| d | Number system base B | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **10** | **9** | **8** | **7** | **6** | **5** | **4** | **3** |
| **1** | 0.301 | 0.315 | 0.333 | 0.356 | 0.387 | 0.431 | 0.500 | 0.631 |
| **2** | 0.176 | 0.185 | 0.195 | 0.208 | 0.226 | 0.252 | 0.292 | 0.369 |
| **3** | 0.125 | 0.131 | 0.138 | 0.148 | 0.161 | 0.179 | 0.208 | |
| **4** | 0.097 | 0.102 | 0.107 | 0.115 | 0.125 | 0.139 | | |
| **5** | 0.079 | 0.083 | 0.088 | 0.094 | 0.102 | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **6** | 0.067 | 0.070 | 0.074 | 0.079 | | | | |
| **7** | 0.058 | 0.061 | 0.064 | | | | | |
| **8** | 0.051 | 0.054 | | | | | | |
| **9** | 0.046 | | | | | | | |

Figure 3. Frequency distribution of first significant digits for various types of number base systems B=3 ,4, 5, ,10




Another property of Benford's Law is that it fulfils the differences between the sorted elements of a set of numbers (either ascending or descending) on condition that starting numbers behave according to Benford Law.

## 4. Benford's Law Conditions

Benford's Law functions best if data possess the following properties:

– **Sufficiently large variability**; the more diversified type of data the better, (length of mp3 tracks does not fall into this category, on the other hand the lengths of all files stored on your computer meet this condition).
– **Lack of determined maximum** or allowed variability limits.
– **Large sample**; the more data the better.
– **Positive distribution asymmetry** (i.e. arithmetic mean is greater than median). In another words, it is better if there are smaller than larger units
– **Random selection of numerous populations** where data are derived.
– **Data result from measurement or count procedures.**
– **Data at transactional level**: for example stock exchange auctions, certificate of current expenses, sales invoice.

Benford's Law does not function in the following situations:
– **Existence of maximum and minimum** value thresholds, for example exam grades (2 to 5), IQ level (50 to 200), age of high school graduates (14 to 18), 400 meter race results, height and weight of persons.
– **Specified permissible value limits** such as the limit for weekly office expenses without management consent (100PLN).
– **Legal and formal restrictions** such as the necessity to pay tax for a sale or purchase above 1000 PLN.
– **Presence of psychological numbers** such as prices at everything below \$1,99"[5].
– **Data obtained from random number generators** i.e. lottery results.
– **Identification data**: numbers of vehicle registration plates, personal ID cards, phone numbers, post codes, product codes, bank account numbers, numbers pertaining to personal ID registration codes valid in Poland (NIP, REGON, ISBN, ISSN, ISMN, IACS).

## 5. Benford's Law Generalizations

Benford's Law can be generalized and subsequently it can be utilized for more sophisticated types of analyses, which employ the following types of tests:
F1 – first significant digit
F2 – first 2 significant digits
F3 – first 3 significant digits
L2 – last 2 significant digits
D2 – exactly second significant digit

Formula 1 indicates the frequency for **first** significant digits (**test F1**). This formula can be generalized for subsequent digits which occur in multi-digit numbers.
For the **second digit** in a sequence (**test D2**) the formula is expressed in the following manner:

$$P(D_2 = d_2) = \sum_{k=1}^{9} \log[1 + \frac{1}{10k + d_2})] \qquad (d_2 = 0,1,2,...,9) \tag{3}$$

Finally, for any given n-th digit combination the formula is expressed accordingly:

[5] This slogan reflects the clients' perceptron that 1.99\$ is substantially lower than 2.00\$ and it will certainly encourage clients to purchase products. According to marketing analysis, such a slogan enables to increase turnover by 10–15%.

(4) $$P(D_1 = d_1; D_2 = d_2; ..., D_n = d_n) = \log\left[1 + \sum_{i=1}^{n} d_i 10^{n-1}\right]$$

From formula (4) we are able to determine the frequency of occurrence for any given two-element digit combination for first 2 significant decimal places **(test F2):**

(5) $$P(D_1 = d_1; D_2 = d_2) = \log\left[1 + \frac{1}{10d_1 + d_2}\right]$$

This operation can be performed for the **first 3** significant decimal places (**test F3**)

(6) $$P(D_1 = d_1; D_2 = d_2; D_3 = d_3) = \log\left[1 + \frac{1}{100d_1 + 10d_2 + d_3}\right]$$

The appearance of specific digits on specific decimal places within the numbers is not dependent on each other. The formula which expresses the conditional frequency of digit occurrence on the second place (d2) under the circumstance that the first place is occupied by digit (d1) is written as follows:

(7) $$P(D_2 = d_2 | D_1 = d_1) = \frac{\log[(d_1 d_2 + 1)/d_1 d_2]}{\log[(d_1 + 1)/d_1]}$$

## 6. Applications of Benford Law

Benford's Law is utilized in different science and real life domains.

**Economic studies**:
- Detection of fraudulent data or unintentional errors in accountancy.
- Detection of tax fraud.
- Analysis of stock exchange data (price and turnover of bonds).
- Analysis of product prices which have been auctioned on the Internet.
- Analysis of the length of time clients feel they are associated with a company which offers services.

– Assessment which pertains to the correctness of compensation estimates in insurance companies.
– Assessment of credibility rates pertaining to fines and penalties ruled in legal proceedings.

**Quantitative research**:
– Testing the correctness of econometric models i.e. forecasting procedure (theoretical data should be subjected to Benford's Law to the same extent as empirical data).
– Optimalization of calculations in solving transport and logistics problems (Euclidean distances between different sites fulfil Benford Law).
– Verifying the correctness of statistical data.

**Technical studies and information technology studies**:
– Designing mass memory for computers[6].
– Being able to distinguish real photography from computer generated graphics.
– Analysis of the capacity of files transferred via Internet, including the time of transfer.
– Benchmarking of digit algorithms.

**Earth sciences**:
– Assessment of clinical efficacy of drugs and medications.
– Analysis of data relevancy pertaining to the emission of toxic pollutants.

It should be pointed out that Benford's Law is not always used in scientific research. Sometimes a scientist will analyze the distribution of significant digits based on data which are relevant and indicate no flaws whatsoever. Subsequently the distribution of digits is performed, but not according to Benford's Law but based on a specific distribution which seems adequate for a particular dataset.
One of the most unique applications of Benford's Law is the analysis of the end of the world prophecy. In their study (1) authors present the frequencies of first significant digits pertaining to the numbers which determine the passage of years from the time when a prophecy is announced until the predicted time of the end of the world. In general the frequencies are consistent with Benford Law; excluding digits: 4, 5 and 7. It should be noted that 7 is the only digit which occurs nearly twice as often in the prophecies (refer to table 3). The fact that sum of numbers doest not add up to 100% in the empirical sequence is due to the unclear nature of prophecies (i.e. different dates for a given prophecy).

[6] Among all number systems the 8 base system enables to save the most space on computer disc, by assuming that we effectively utilize Benford's Law properties. In addition to that, the construction of logarythmic computers which utilize information, comprising a greater frequency of numbers which began from small digits, enables to accelarate the time of variable comma calculations.

Table 3. Conformity of first significant digits distribution for dates of end of the world prediction according to Benford Law

| **I digit** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
|---|---|---|---|---|---|---|---|---|---|
| Benford | 30.1 | 17.6 | 12.5 | 9.7 | 7.9 | 6.7 | **5.8** | 5.1 | 4.6 |
| Prophecies | 28.8 | 16.7 | 12.1 | 6.1 | 3.0 | 6.1 | **10.6** | 4.5 | 4.5 |
| Difference | 1.3 | 0.9 | 0.4 | 3.6 | 4.9 | 0.6 | **-4.8** | 0.6 | 0.1 |

Another interesting example is that the number of hits in web search engines agrees with Benford law. Usually we have to type in, as key words, phrases containing several numbers starting with 1,2,3,…,9; for example: *1ddmmyy, 2ddmmyy, 3ddmmyy*,…, where *ddmmyy* denotes date (day, month, year). It turns out that we will always obtain monotonic sequences which possess descending frequencies.

## 7. Proposition to Create a Portal that would Analyze Data in Terms of Conformity with Benford Law

The discussion presented in this paper indicates that Benford's Law may be applied only in certain situations; it should be noted that the strength of Benford's Law impact (measured by the degree of conformity with the Benford formula) is generally diversified. The number of analyses that have been performed so far regarding Benford's Law (even if its a large number) does not enable to conclude that this law is common in nature.

Therefore it is recommended to create a portal which would operate according to Web 2.0 principle. The goal of this portal would be to gather information pertaining to the results of conformity analysis with Benford's Law based on various sets of data. This portal should include the following components, and subsequently enable the accomplishment of the formulas mentioned below:

1. Articles contaning a thorough description of the essence, properties and conditions of Benford's Law (Wikipedia).
2. On-line applications which enable performing statistical analysis on any given set of data, according to a standardized calculation scheme and range of parameters.
3. Possibility to post the results of such analysis on the portal including brief commentary.
4. Possibility to comment on the results that have been posted by other users (via blog).
5. Ranking of posted analyses from the standpoint of their uniqueness.

6. Generating classifications of analyses based on their context related field of applications.

The most important component of this portal is its analytical application, which would include various types of starting data formats as well as standard set of analysis results (chart type, result table, measures of conformity i.e. chi square test, approximation errors etc.)
This particular portal could be utilized as a didactical aid, for example during a seminar on the methods of quantitative data. One of the forms for completing such a course might be the requirement to perform a specific analysis procedure on a given set of data (obtained from real time database) by employing software tools available on this portal. Subsequently, a student would be obliged to post a description of obtained results.

**Abstract**

The paper presents Benford's Law, also called the first digit law, stating that in lists of numbers from many real-life sources of data the leading digit is distributed in a specific non-uniform way. The essence, properties, constrains as well as examples of Benford's Law practical uses have been shown. It has been proposed to build a Web 2.0 site (vertical portal) where results of different data set analyses could be reported and presented in a unified way, in the context of their agreement with Benford's Law.

**Literature**

Benford F., *The law of anomalous numbers*, „Proceedings of the American Philosophical Society" 1938, No. 78.

Frey B., *75 sposobów na statystykę. Jak zmierzyć świat i wygrać z prawdopodobieństwem*, Helion, Gliwice 2007.

Hill T.P., *Benford's law, Encyklopedia of Mathematics*. *Supplement*, „Kluwer" 1997, No. 1.

Hill T.P., *A statistical derivation of the significant digit law,* „Statistical Science" 1996, No. 10.

Karpiński M., *Prawo Benforda*, „Matematyka w Szkole" 2008, No. 31.

Newcomb S., *Note on the frequency of use of the different digits in natural numbers,* „American Journal of Mathematics" 1881, No. 4.

Nigrini M., *A taxpayer compliance application of Benford's law,* „Journal of the American Taxation Association" 1996, No. 18.

Paulos J.A., *Matematyk gra na giełdzie*, Wydawnictwo CDW, Azymut–Book 2007.

Weisstein E.W., *Benford's law,* http://mathworld.wolfram.com/BenfordsLaw.html.

Wołowik P., *Prawo Benforda – testowanie i weryfikacja poprawności danych pomiarowych***,** „Przegląd Telekomunikacyjny – Wiadomości Telekomunikacyjne" 2005, No. 11.

**Web sites used**

http://en.wikipedia.org/wiki/Benford%27s_law.
http://mathworld.wolfram.com/BenfordsLaw.html.
http://www.intuitor.com/statistics/Benford%27s%20Law.html.
http://www.mathpages.com/home/kmath302/kmath302.htm.
www.rexswain.com/benford.html.