MÄLARDALEN UNIVERSITY
SWEDEN

School of Education, Culture and Communication
Division of Mathematics and Physics

BACHELOR'S DEGREE PROJECT IN MATHEMATICS

Benford's Law: Analysis of the trustworthiness of COVID-19 reporting in the context of different political regimes

by

Nikolaos Giannakis, Leonid Burlac

MAA322 — Examensarbete i matematik för kandidatexamen

DIVISION OF MATHEMATICS AND PHYSICS
MÄLARDALEN UNIVERSITY
SE-721 23 VÄSTERÅS, SWEDEN

MÄLARDALEN UNIVERSITY
SWEDEN

School of Education, Culture and Communication
Division of Mathematics and Physics

MAA322 — Bachelor's Degree Project in Mathematics

Date of presentation:
2nd of June 2021

Project name:
Benford's Law: Analysis of the trustworthiness of COVID-19 reporting in the context of different political regimes

Authors:
Nikolaos Giannakis, Leonid Burlac

Version:
4th June 2021

Supervisor(s):
Milica Rančić

Reviewer:
Hang Zettervall

Examiner:
Jean-Paul Murara

Comprising:
15 ECTS credits

Abstract

In order for governments and demographers to, among other things, design policies and pension plans, as well as for insurance companies to offer policies that serve general public, having reliable mortality data plays a crucial role. The academic world works actively in developing tools (models and methods) that can, based on collected mortality data, forecast future death rates in the observed population. Obviously, to be able to rely on the predicated data one needs a reliable source of existing mortality data. In the light of the ongoing COVID-19 pandemic, reliability of certain death-case reporting has been questioned. In this thesis, the Benford's Law is used to evaluate how well countries with authoritarian regimes (Azerbaijan, Belarus), and with democratic regimes (Greece, Serbia, Sweden), report their COVID-19 cases to the worldwide public. Statistical tests such as the Chi-squared test, mean absolute deviation, and the distribution distance were used to obtain the results needed to form our conclusions. During our testing, we found that countries with democratic regimes do conform better to the Benford's law than the authoritarian ones.

Acknowledgements

Firstly, we would like to express our gratitude towards our supervisor Milica Rančić for her guidance through this paper, all the great advises and useful content she has provided us, and her patience that helped us to conduct this study. We would also like to thank Mälardalen University for all the knowledge that we have obtained through our time of studying.

List of Abbreviations

BL	Benford's Law
r.v.	Random variable
WHO	World Health Organization
PDF	Probability density function
MAD	Mean absolute deviation
df	Degrees of freedom

List of Symbols

∞	infinity
$\in$	belongs in
$\bigcup$	arbitrary union
$\cup$	union
log	logarithm base 10
$\mathscr{M}$	significand σ-algebra
$\mathbb{N}$	set of natural numbers
$\mathbb{R}^+$	set of positive real numbers
a	significance level
B	Borel set
b	BL distribution for each digit
b_d	Benford's Law distribution
d	$d = (1, 2, ..., 9)$ where 1,2,...,9 are the first digits
d^*	distribution distance test statistic
$E(X)$	expected value of random variable X
e	the number e or Euler's number
$f(x)$	density function of X
H_0	null hypothesis
H_a	alternative hypothesis
h_d	observed different frequencies
k	degrees of freedom
N	number of data points
N_d	number of observations of the integer d
n	number of observations / sample size

$S(x)$	significand
s	scalar
$V(X)$	variance of random variable X
X	random variable
α	shape parameter
β	scale parameter
$\Gamma(\alpha)$	gamma function
θ	inverse scale parameter
μ	mean
$\sum$	summation or sum
σ^2	variance
χ^2	Chi-square test

Contents

Acknowledgements **2**

List of Abbreviations **3**

List of Symbols **3**

1 Introduction **8**
1.1 Overview 8
1.2 Aim and Purpose 9
1.3 Methodology 9

2 Theoretical Consideration **11**
2.1 Generalization (Statement of Benford's Law) 11
2.1.1 The distribution of the first significant digit 11
2.1.2 The distribution of significant digits 12
2.1.3 Mantissa (significand) distribution 13
2.1.4 The significand σ-algebra 15
2.1.5 Scale and base invariance 16
2.2 Proof 16
2.3 Limitations 19
2.4 Distributions 19
2.4.1 The Gamma Distribution 20
2.4.2 Chi-square Distribution 21
2.5 Statistical tests 23
2.5.1 The Chi-square test of goodness-of-fit 23
2.5.2 Distribution distance 24
2.5.3 Mean Absolute Deviation 25

3 Methodology **27**
3.1 The data 27
3.2 Testing thesis method 28
3.3 Data Analysis 29

4 Results **31**

5 Conclusion **35**
5.1 Thesis summary 35
5.2 Future work 36

6 Reflection of objectives in the thesis **37**
6.1 Objective 1: Knowledge and understanding 37
6.2 Objective 2: Methodological knowledge 37
6.3 Objective 3: Critically and Systematically Integrate Knowledge 37
6.4 Objective 4: Independently and Creatively Identify and Carry out Advanced Tasks 38
6.5 Objective 5: Present and Discuss Conclusions and Knowledge 38
6.6 Objective 6: Scientific, Social and Ethical Aspects 38

Bibliography **39**

Chapter 1

Introduction

1.1 Overview

Nowadays, everything is related to enormous amounts of data. Satellites provide daily information greater than the entire Kungliga biblioteket (The National Library of Sweden) meaning that the researchers need to efficiently and quickly analyze these sets of data. Consequently, individuals are interested in patterns of data. Benford's Law (BL) is one of these applications that analyze data patterns and has to do with how frequently the leading digits or first digits appear. The concept of scientific notation was introduced which is that: a nonzero number y can be written as $S(y) * 10^k$, where $S(y) \in [1, 10)$ is the significand and k is an integer. This integer part is the leading digit or the first digit [21].

Although, the law holds Benford's name, in reality he was not the first to observe such a leading digit distribution. Simon Newcomb (1835-1909) who was an astronomer-mathematician noticed this behaviour almost five decades before Benford [21]. One of Newcomb's short articles indicates that digits do not occur with the same probability and that the most frequent occurring first digit is integer 1 whereas 9 is the digit with occuring the least. Furthermore, the paper also notes that is crucial to not select natural numbers at random but to choose two specific ones and then find the probability of the first significant digit n with the help of their ratio [22].

Frank Benford was a physicist in the Research Laboratory of the General Electric Company in Schenectady, NY, USA and his work there was most related with optics. The notable law is also known under the name of "The Law of Anomalous Numbers". Moreover, he was the one to study the distribution of twenty different sets of data, such as area, population, rivers, newspapers etc. and check this kind of leading digit behaviour. Noteworthy is one important finding in his study that indicates that while individual sets may not satisfy BL, connecting different data of sets forms a sequence which seems to behave similar to the corresponding law [21].

Benford's Law arises in a variety of disciplines and few of them take place afterwards. In electrical engineering by using lightning data in order to check if the data follows the BL distribution [19]. The document is using data taken from the European Cooperation for Lightning Detection. It then applies a Chi-square goodness-of-fit test in order to examine if

the two considered data sets named Lightning Peak current and Inderstroke interval follow the BL distribution which in fact they do. In addition, the law can be found, in biological sciences according to the paper [9]. Four hundred and nine Microcystis *aeruginosa* colonies were collected from different locations in Andalusia, Spain and their number of cells were analyzed by using Chi-square goodness-of-fit test with eight degrees of freedom. The result gives that the number of cells of the certain cyanobacterium follow the BL distribution. Furthermore, in social sciences through the study of the paper [14] , where the study analyzes the data of five major social networks. Four of them followed the BL distribution, whereas one did not due to a feature of that platform that was able to change the individuals' behavior. Finally, in accounting according to [3], where the authors indicate that Benford analysis can be used in order to examine if there are patterns in huge number of data that show clues of manipulation.

After introducing some general background about this law, BL can be defined as follows:

$$N_d = Nlog_{10}(1 + \frac{1}{d})$$

where N is the number of data points and N_d is the number of observations of the integer $d = (1, 2, ..., 9)$. Often the law fails to be satisfied if there is human manipulation or flaws in the given data [16]. Therefore, BL has been used to identify fraudulent or manipulated data of different nature. BL can be potentially used in order to examine if a specific country has given false or manipulated COVID-19 data presented to the public. Since the spread of virus exhibits exponential growth and changes in terms of magnitude, the law can be applied to these types of data. It has to do with the fact that Benford distribution of the leading digits appears naturally for such exponential events with changes in the magnitude [6].

1.2 Aim and Purpose

This study is carried out with the purpose of analyzing the BL from the scratch, as well as its derivation, generalizations and limitations. Deficiencies and improvements will also be discussed. Finally, this thesis attempts to further clarify and add knowledge to previous research by implementing the law on five countries' COVID-19 data and examine if this specific data is trustworthy. The study will be held in the context of Greece, Serbia, Sweden, Belarus and Azerbaijan. This way the study will provide important information to the public, as well as government and demographic planning bodies, insurance companies etc, who greatly rely on trustworthiness of this data in their work. Moreover, the academic community will be enriched with additional statistical experiments. Our hope is that, during the course of the project, future research ideas will arise which can be suggested to the research community for further investigation.

1.3 Methodology

It is crucial to study and understand BL in order to attain better background knowledge about the topic. In addition, Chapter 2 will consist of the proof, generalizations and limitations

of the law. Further, we will use quantitative secondary data taken from *Center for Systems Science and Engineering of John Hopkin's University* [5] or any other trustworthy source for the COVID-19 cases. The data will be taken from countries that, according to the authors' knowledge, are not investigated in any other research. We intend to use a programming language such as R and MatLab in order to show any false reporting from the chosen countries, by illustrating our results using graphs and tables. Furthermore, we may use programming codes that are already mentioned in previous research.

Chapter 2

Theoretical Consideration

2.1 Generalization (Statement of Benford's Law)

In order to better understand the Benford's law and the way it performs, we have written down relevant definitions and such that are based on [13], [24], [21], and [25].

2.1.1 The distribution of the first significant digit

Definition 2.1.1 (Benford's Law for the first significant digit)**.** We say the data set satisfies Benford's Law for the Leading digit if the probability of observing a first digit of d is *approximately*

$$P(D_1 = d_1) = \log \frac{d+1}{d}, \tag{2.1}$$

for d = 1,2,...,9.

It is hard to say what approximately might mean in this case. By conducting a statistical test, such as the most commonly used in this case Chi-square, it often rejects the null hypothesis with large data sets if there is a small deviation from the distribution. Thus, besides the results of the null hypothesis testing, we shall consider a good visual fit to describe the word "*approximately*" in our study. In addition, for a better understanding, Fig. 2.1 reveals the ratio of each of the 9 digits from 1 to 9 that follow the Benford's law.

Despite determining the probability of the first significant digit, it is possible to find the probability of the entire significand, meaning that we can find the probability of observing a significand between 1 and 2, or between e and π. This is referred as the Strong Benford's Law.

Definition 2.1.2 (Strong Benford's Law for the Leading Digits)**.** The data satisfying the Strong Benford's Law would be if the probability of observing a significand in [1,s) is $\log s$.

The under-performance of this law for certain types of data sets has lead to an establishment of certain criteria that should be met so that the data would obey the law. These include:

- Relatively uniform data span over several magnitudes.

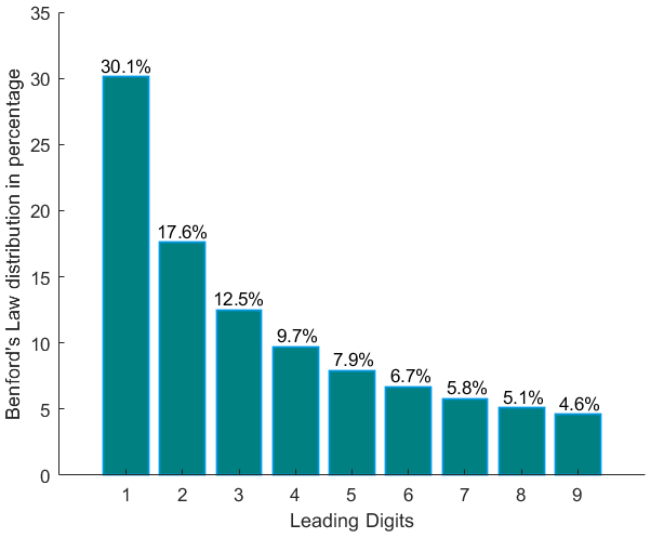

Figure 2.1: Benford's Law for the first signifinact digit

- The mean should be greater than the median, and have a positive skew.

- The data under testing should be of a natural occurrence, the result of multiplicative variation, and not modified by any human involvement.

2.1.2 The distribution of significant digits

Definition 2.1.3 (The general significant-digit law)**.** For all positive integers c, all $d_1 \in \{1,2,...,9\}$ and all $d_j \in \{0,1,...,9\}$ where $j = 2, ..., c$, it follows that

$$P(D_1 = d_1, ..., D_c = d_c) = \log \left[1 + (\sum_{i=1}^{c} d_i \times 10^{c-i})^{-1} \right]. \tag{2.2}$$

By using the above definition, the probabilities for the first and for the second significant digits are presented in the Table 2.1.

Table 2.1: Probabilities for the first and second significant digits under Benford's Law

Digit	0	1	2	3	4	5	6	7	8	9
First	-	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%
Second	12.0%	11.4%	10.9%	10.4%	10.0%	9.7%	9.3%	9.0%	8.8%	8.5%

The author of [24] mentions that the significant digits are dependent, meaning that the unconditional probabilities of the j^{th} significant digits differ from the conditional probabilities when the $j^{th} - 1$ significant digit is given.

2.1.3 Mantissa (significand) distribution

From Definition 2.1.3 we can generalize the mantissa distribution as follows [13]:

Lemma 2.1.1. The logarithmic density in Definition 2.1.3 can be generalized in a continuous way for the mantissa M in the following form:

$$P(M \leq m) = \log(m), \tag{2.3}$$

where $m \in [1,10)$.

Proof.

- First, we consider the case where $m = d_1$, i.e. m has only one significant digit:

$$P(M \leq m) = \begin{cases} 0 \text{ if } d_1 \leq 1 \ (P(M = 1) = 0) \text{ in a continuous case)} \\ P(D_1 \leq d_1 - 1) = \log(d_1) = \log(m) \text{ if } 1 < d_1 < 10. \end{cases} \tag{2.4}$$

- Then we suppose $m = d_1, d_2, ..., d_c$ in decimal notation, i.e. $m = \sum_{i-1}^{c} 10^{-(i-1)} d_i$, with $d_1 > 1, d_2 > 0, ..., d_c > 0$:

$$\begin{aligned} P(M \leq m) &= P(D_1 \leq d_1 - 1) + \\ &\quad + P(D_1 = d_1, D_2 \leq d_2 - 1) + \\ &\quad + ... + \\ &\quad + P(D_1 = d_1, D_2 = d_2, ..., D_{c-1} = d_{c-1}, D_c \leq d_c - 1) \\ \\ &= P(D_1 \leq d_1 - 1) + \\ &\quad + \sum_{0 \leq d_2' \leq d_2 - 1} P(D_1 = d_1, D_2 = d_2') + \\ &\quad + ... + \\ &\quad + \sum_{0 \leq d_c' \leq d_c - 1} P(D_1 = d_1, D_2 = d_2 - 1, ..., D_c = d_c') \\ \\ &= \log(d_1) + \\ &\quad + \sum_{0 \leq d_2' \leq d_2 - 1} \log\left(1 + \frac{1}{10 d_1 + d_2'}\right) \\ &\quad + ... + \\ &\quad + \sum_{0 \leq d_c' \leq d_c - 1} \log\left(1 + \left(\sum_{i=1}^{c-1} 10^{c-i} d_i + d_c'\right)^{-1}\right). \end{aligned} \tag{2.5}$$

By analogy with the derivation of the first digit distribution:

$$\begin{aligned} P(D \leq d) &= \sum_{1 \leq d' \leq d} P(D = d') = \sum_{1 \leq d' \leq d} \log\left(1 + \frac{1}{d'}\right) \\ &= \log\left(\prod_{1 \leq d' \leq d} \left(1 + \frac{1}{d'}\right)\right) \\ &= \log\left(\left(1 + \frac{1}{1}\right)\left(1 + \frac{1}{2}\right) ... \left(1\frac{1}{d}\right)\right) \\ &= \log\left(\frac{2}{1} \times \frac{3}{2} \times ... \times \frac{d+1}{d}\right) \\ &= \log(d+1), \end{aligned} \tag{2.6}$$

where $d \in \{1, ..., 9\}$, we find:

$$\begin{aligned} P(M \leq m) &= \log(d_1) + \log\left(\frac{10d_1 + d_2}{10d_1}\right) + ... + \log\left(\frac{\sum_{i=1}^{c} 10^{c-1} d_i}{\sum_{i=1}^{c-1} 10^{c-i} d_i}\right) \\ &= \log\left(\frac{\sum_{i=1}^{c} 10^{c-i} d_i}{10^{c-1}}\right) \\ &= \log\left(\sum_{i=1}^{c} 10^{-(i-1)} d_i\right) \\ &= \log(m). \end{aligned} \tag{2.7}$$

- In case that any of the d_j's ($j > 1$) are null, the above still holds.
  For example, if $m = d_1 d_2 ... d_{j-1} 0 d_{j+1} ... d_c$, then there is no j^{th} term in the sum, which leads to the followng expression:

$$\begin{aligned} P(M \leq m) &= \log(d_1) + ... \\ &\quad + \log\left(\frac{\sum_{i=1}^{j-i} 10^{j-1-i} d_i}{\sum_{i=1}^{j-2} 10^{j-1-i} d_i}\right) + 0 + \log\left(\frac{\sum_{i=1}^{j+1} 10^{j+1-i} d_i}{\sum_{i=1}^{j} 10^{j+1-i} d_i}\right) + ... \\ &\quad + \log\left(\frac{\sum_{i=1}^{c} 10^{c-i} d_i}{\sum_{i=1}^{c-1} 10^{c-i} d_i}\right) \\ &= \log\left(\frac{1}{10^{j-2}} \times \frac{\sum_{i=1}^{j-1} 10^{j-1-i} d_i}{\sum_{i=1}^{j} 10^{j+1-i} d_i} \times \frac{1}{10^{c-j-1}} \times \sum_{i=1}^{c} 10^{c-i} d_i\right) \\ &= \log\left(\frac{1}{10^{j-2}} \times \frac{1}{10^2} \times \frac{1}{10^{c-j-1}} \times \sum_{i=1}^{c} 10^{c-i} d_i\right) \\ &= \log(m). \end{aligned} \tag{2.8}$$

  This can be easiliy extended to the case where several d_j's are null or if $d_1 = 1$.

□

2.1.4 The significand σ-algebra

It is noticeable that the definitions of the significant digit's laws are probabilities, thus, it is important to assign the right probability space, hence the correct σ-algebra. Here, [21] and [24] define it as follows:

Definition 2.1.4. The significand σ-algebra S, denoted by $\mathscr{M}$ and will be called the (decimal) *mantissa σ-algebra*, is the σ-algebra on $\mathbb{R}^+$ generated by the significand function S, i.e., $S = \mathbb{R}^+ \cap \sigma(S)$. It is a subfield of Borels defined by:

$$\mathscr{M} = \bigcup_{n=-\infty}^{\infty} B \times 10^n \tag{2.9}$$

for some Borel $B \subseteq [1, 10)$.

Lemma 2.1.2. Main properties of the mantissa algebra are:

i. Every non-empty set in $\mathscr{M}$ is infinite with accumulation points at 0 and $+\infty$,

ii. $\mathscr{M}$ is closed under scalar multiplication ($s > 0, S \in \mathscr{M} \Rightarrow sS \in \mathscr{M}$),

iii. $\mathscr{M}$ is closed under integral roots ($m \in \mathbb{N}, S \in \mathscr{M} \Rightarrow S^{1/m} \in \mathscr{M}$), but not powers,

iv. $\mathscr{M}$ is self-similar in the sense that if $S \in \mathscr{M}$, then $10^m S = S$ for every integer m.

While properties i, ii, and iv follow easily the definition [24], a closer inspection to the property iii can be done.

Proof. Proof of property iii
The square root of a set in $\mathscr{M}$ may consist of a few parts, and the same goes for higher roots. For instance, if

$$S = \{D_1 = 1\} = \bigcup_{n=-\infty}^{\infty} [1, 2) \times 10^n, \tag{2.10}$$

then

$$S^{1/2} = \bigcup_{n=-\infty}^{\infty} [1, \sqrt{2}) \times 10^n \cup \bigcup_{n=-\infty}^{\infty} [\sqrt{10}, \sqrt{20}) \in \mathscr{M}, \tag{2.11}$$

but

$$S^2 = \bigcup_{n=-\infty}^{\infty} [1, 4) \times 10^{2n} \notin \mathscr{M}, \tag{2.12}$$

because of the great gaps that prevent writing it down in terms of $\{D_1, D_2, ...\}$. □

From the above properties it is worth mentioning that ii is key to the hypothesis of scale invariance and iv is key to the base hypothesis.

2.1.5 Scale and base invariance

Considering the "universality" of the BL, one of the first hypothesis that one may think of is it's scale invariance. The idea that natural data sets would follow the law independent of the chosen unit system means that converting the data by multiplying it with whatever constant will not change the probability measures. Furthermore, it is of interest if the BL is affected by the change of the base, meaning that if an observed data set is in base 10, then the BL would be observed even if the base would be changed. In both of his articles, [24] and [25], the author explains in more detail the theorems behind these properties, from where we can write down the following definitions for each case.

Definition 2.1.5. A probability measure P on $(\mathbb{R}^+, \mathscr{M})$ is *scale invariant* if $P(S) = P(sS)$ for all $s > 0$ and all $S \in \mathscr{M}$.

Definition 2.1.6. A probability measure P on $(\mathbb{R}^+, \mathscr{M})$ is *base invariant* if $P(S) = P(S^{1/n})$ for all positive integers n and all $S \in \mathscr{M}$.

However, there still exist questions concerning scale invariance, such as Furstenberg's 25-year-old conjecture that the uniform distribution on $[0, 1)$ is the only atomless probability distribution invariant under both $2x \pmod 1$ and $3x \pmod 1$ [24].

2.2 Proof

The successfulness of the BL was quite a mystery for many years as it was unclear whether the law was relevant because of some sort of mechanism present in the nature or that it was a result of human system of numbers. This has changed with the general derivation of the law from application of the Laplace transform, where the law is derived in its strict form that is composed of the Benford term that explains the generality of the law, and an error term that leads to deviations from the law. We will present the proof that the authors in [17] have shown, which shows to be very neat and understandable. Although the authors have derived a proof for all the significant digits, we will present the proof for the first digit only as it is our main point of interest in this thesis, and will leave the reference for the reader.

Let $F(x)$ be our probability density function on the set of all real positive numbers $\mathbb{R}^+$. (Note that we are using F instead of the lower case as used in Laplace). It might happen that x could be a negative number as well, but this could be fixed by taking the absolute value of x and using it in the probability density function, and thus keep the results unchanged.

The probability P_d on the decimal system whose value is d is the sum of the probabilities on the interval $[d \cdot 10^n, (d+1) \cdot 10^n)$ for all integers n, thus P_d can be written as:

$$P_d = \sum_{n=-\infty}^{\infty} \int_{d \cdot 10^n}^{(d+1) \cdot 10^n} F(x)dx, \tag{2.13}$$

which can be rewritten as:

$$P_d = \int_0^{\infty} F(x) g_d(x) dx, \tag{2.14}$$

where $g_d(x)$ will be a new density function whose role will be explained from now on. By adopting the Heaviside step function,

$$\eta(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0, \end{cases} \tag{2.15}$$

we can write $g_d(x)$ as

$$g_d(x) = \sum_{n=-\infty}^{\infty} [\eta(x - d \cdot 10^n) - \eta(x - (d+1) \cdot 10^n)]. \tag{2.16}$$

We can explain from above that in the decimal system, numbers favor to smaller first digits, opposed to the thought that each of the numbers from 0 to 9 have the same probabilities. Figure 2.2 will be more helpful to understand why this happens. The density functions $g_1(x)$ and $g_2(x)$ are presented on the interval [1,30).

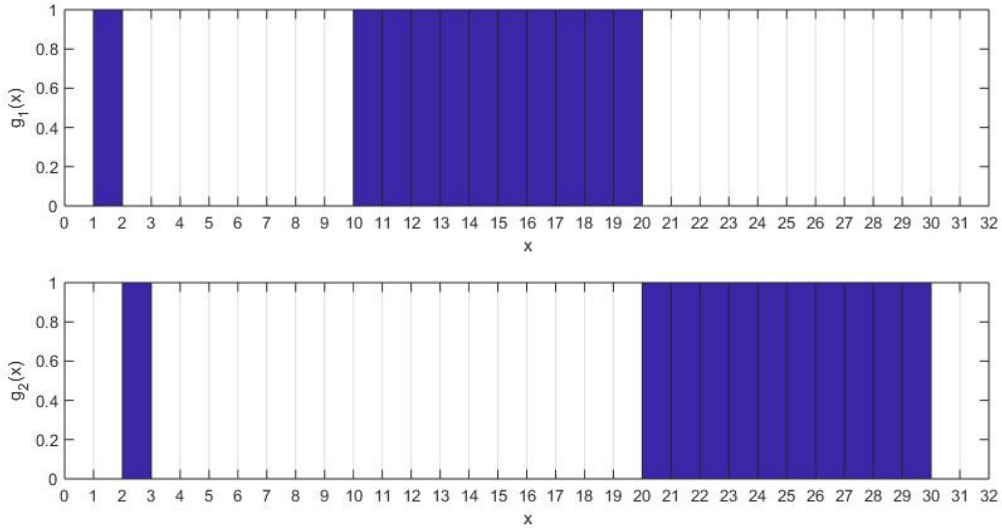

Figure 2.2: Images of $g_1(x)$ and $g_2(x)$ functions showing their distribution

We now prove that if the PDF has an inverse Laplace, it satisfies BL. Let $f(t)$ be the inverse Laplace transform of $F(x)$, and $G(t)$ be the Laplace transform of $g(x)$, i.e.

$$F(x) = \int_0^{\infty} f(t) d^{-tx} dt, \tag{2.17}$$

$$G(t) = \int_0^{\infty} g(x) d^{-tx} dx. \tag{2.18}$$

The Laplace transform's properties are the following:

$$\begin{aligned}\int_0^\infty F(x)g(x)dx &= \int_0^\infty dx g(x) \int_0^\infty f(t)e^{-tx}dt \\ &= \int_0^\infty dt f(t) \int_0^\infty g(x)e^{-tx}dx \\ &= \int_0^\infty f(t)G(t)dt,\end{aligned} \tag{2.19}$$

meaning that Laplace may act on either f or g with the above integral without changing the results.
To calculate the left-hand side, for convenience the righ-hand side will be calculated. Beginning with Laplace transform of $g_d(x)$, it gives:

$$\begin{aligned}G_d(t) &= \int_0^\infty g_d(x)e^{-tx}dx \\ &= \sum_{n=-\infty}^{\infty} \int_{d\cdot 10^n}^{(d+1)\cdot 10^n} e^{-tx}dx \\ &= \frac{1}{t}\sum_{n=-\infty}^{\infty}(e^{-td\cdot 10^n} - e^{-t(d+1)\cdot 10^n}),\end{aligned} \tag{2.20}$$

which can be treated as a function of variables d and t. Although d is defined on the decimal digit set 1,2,...,9, it can be extended to the whole real axis and thus $G_d(t)$ is continuous for both d and t. To evaluate $G_d(t)$, the partial derivative will be calculated with respect to d, and integrate the partial derivative that results in

$$\begin{aligned}\frac{\partial G_d(t)}{\partial d} &= \sum_{n=-\infty}^{\infty}(-10^n e^{-td\cdot 10^n} + 10^n e^{-t(d+1)\cdot 10^n}) \\ &\approx \int_{-\infty}^{\infty}(-10^x e^{-td\cdot 10^x} + 10^x e^{-t(d+1)\cdot 10^x})dx \\ &= \frac{1}{\ln 10}\int_0^\infty(-e^{-tdy} + e^{-t(d+1)y})dy \\ &= \frac{1}{\ln 10}\left(-\frac{1}{td} + \frac{1}{t(d+1)}\right).\end{aligned} \tag{2.21}$$

Because $G_d(t) \to 0$ when $d \to \infty$,

$$G_d(t) \approx \frac{1}{t}\log_{10}(1 + \frac{1}{d}) \tag{2.22}$$

and thus:

$$\begin{aligned} P_d &= \int_0^\infty F(x) g_d(x) dx \\ &= \int_0^\infty G_d(t) f(t) dt \\ &\approx \int_0^\infty \frac{f(t)}{t} \log_{10}(1 + \frac{1}{d}) dt \\ &= \log_{10}(1 + \frac{1}{d}) \int_0^\infty \frac{f(t)}{t} dt \\ &= \log_{10}(1 + \frac{1}{d}), \end{aligned} \tag{2.23}$$

where the following normalization condition of $f(t)$ has been used:

$$\begin{aligned} 1 &= \int_0^\infty F(x) dx \\ &= \int_0^\infty dx \int_0^\infty f(t) e^{-tx} dt \\ &= \int_0^\infty dt f(t) \int_0^\infty e^{-tx} dx \\ &= \int_0^\infty \frac{f(t)}{t} dt. \end{aligned} \tag{2.24}$$

2.3 Limitations

Although Bernford's law is largely adopted for data checking in various fields, in some cases though this method performs poorly in indicating a deviation that can suspect a fraud in the data. An obvious limitation would be a really small data set. The law can be observed only over a big collection of data. Furthermore, the authors of [3] explained the most notable limitations for this method. The method can detect deviation in proportion in case that some data has either been added or removed, which in result will break the chain of natural occurrence. However, if the data has not been added at all, it cannot violate the occurrence, and here this method shows a significant downside. Another case of poor performance of this method is when the data has a limited magnitude in it's values, for instance if an input of data requires the number to be within a specific region (e.g. from 20 to 500). The leading digits in this case would not follow the law merely because the data will omit lower or higher entries, breaking the natural proportion. Prices that are assigned by humans are not compatible with this law either, as well as assigned numbers to e.g. accounts, transactions etc., and firm specific numbers.

2.4 Distributions

Before reaching to Chi-square test couple of concepts are crucial to be indicated in this thesis for a better understanding. In order to explain Chi-square test there is a need for Chi-

square distribution to be stated. However, Chi-square distribution is a special case of Gamma distribution which means that this concept should also be indicated [20]. Finally, in the end some subsections may have a small numerical example representing how the aforementioned concepts can be applied.

2.4.1 The Gamma Distribution

Definition 2.4.1 (Gamma Distribution)**.** A random variable (r.v.) X will have a Gamma distribution with the following parameters $\alpha > 0$ and $\beta > 0$ if and only if the density function of X is

$$f(x) = \begin{cases} \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}, & \text{for } 0 \leq x < \infty, \\ 0 & \text{x<0} \end{cases} \tag{2.25}$$

where

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}dx. \tag{2.26}$$

The integral is the well known gamma function. Further, Fig. 2.3 indicates that for different values of α and β the shape of the gamma density changes according to α. Consequently, that is parameter α is called *shape parameter*. Whilst, parameter β is known under the name *scale parameter* since when someone multiplies a gamma-distributed r.v. by a positive constant the result will be again an r.v. following a gamma distribution. The only difference is that β will be revised but α stays the same. In addition, there is the inverse scale parameter or rate parameter $\theta = \frac{1}{\beta}$ which is going to help us simplify the density function later.

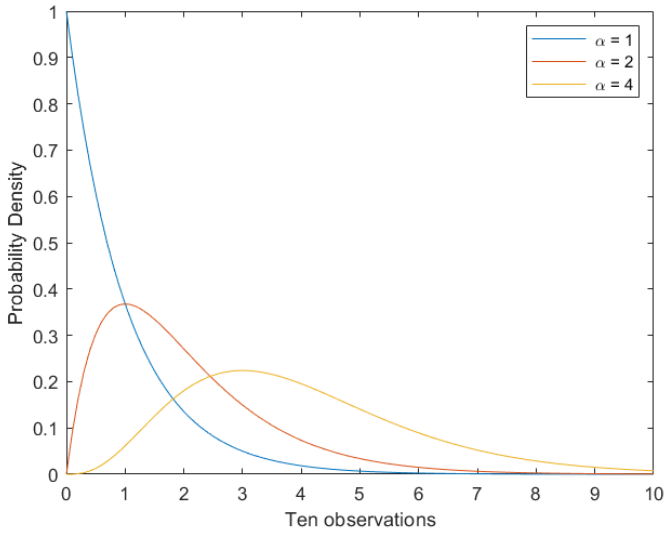

Figure 2.3: Γ density functions with different α

The following example sheds light upon on how to use the concept of Gamma Distribution in the real world in order to predict a certain probability.

Example 2.4.1. *The magnitude of earthquakes that were recorded in a region of Greece follows a gamma distribution with $\alpha = 0.6$ and $\beta = 2.3$. What is the probability that the magnitude of an earthquake striking that region will exceed the 4.5 on the Richter scale?*

Let X be the magnitude of an earthquake which strikes in a region measured by the Richter scale $X \sim \Gamma(\alpha, \beta) = \Gamma(0.6, 2.3)$ which means that X follows a Γ distribution with the corresponding α and β.

Here it is enough to use an applet or the table of Γ distribution and find the corresponding probability which is $P(X > 4.5) \approx 0.06305$. We decided to use a software in order to find the probability that the magnitude will be greater than 4.5 on the Richter scale [8].

2.4.2 Chi-square Distribution

In order to prove the latter Chi-square distribution's definition, it is essential to refer to a gamma distribution's theorem.

Theorem 2.4.1. *If X has a gamma distribution with parameters α and β, then*

$$\mu = E(X) = \alpha\beta \text{ and } \sigma^2 = V(X) = \alpha\beta^2, \tag{2.27}$$

where μ is the mean, $E(X)$ is the expected value of the r.v. X, σ^2 and $V(X)$ are the variance of the r.v X.

Proof. Now it is important to prove these two equalities. It is known that the expected value is equal to

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx. \tag{2.28}$$

It is already known that $f(x)$ is equal to eq. (2.25). Thus,

$$E(X) = \int_{0}^{\infty} x \left(\frac{x^{\alpha-1} e^{\frac{-x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)} \right) dx. \tag{2.29}$$

The integral's limits are different due to the two cases in eq. (2.25). Additionally, we know that the gamma density function integrates to 1 and we need this mathematical concept in order to further proceed with our proof of expected value. Consequently, substituting the inverse scale parameter $\theta = \frac{1}{\beta}$ follows that:

$$\int_{-\infty}^{\infty} f(x) dx = \int_{0}^{\infty} \frac{\theta^{\alpha} x^{\alpha-1} e^{-x\theta}}{\Gamma(\alpha)} dx = \frac{1}{\Gamma(\alpha)} \theta^{\alpha} \int_{0}^{\infty} x^{\alpha-1} e^{-x\theta} dx. \tag{2.30}$$

For now leave the constant part out and analyze the integral part. Let $t = \theta x$, then $dx = \frac{1}{\theta} dt$ and

$$\frac{1}{\theta} \int_{0}^{\infty} \left(\frac{t}{\theta} \right)^{\alpha-1} e^{-t} dt = \frac{1}{\theta^{\alpha}} \int_{0}^{\infty} t^{\alpha-1} e^{-t} dt. \tag{2.31}$$

The part inside the integral looks familiar and it is actually $\Gamma(\alpha)$ according to eq. (2.26),

$$\frac{1}{\theta^{\alpha}}\Gamma(\alpha). \tag{2.32}$$

In addition refer back and get the constant that was left out before. Thus,

$$\frac{\theta^{\alpha}}{\Gamma(\alpha)}\frac{\Gamma(\alpha)}{\theta^{\alpha}} = 1. \tag{2.33}$$

Since we proved that the gamma density function is equal to 1, it can be used to prove Theorem 2.4.1. Thus,

$$\int_0^{\infty} \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}dx = 1. \tag{2.34}$$

Consequently,

$$\int_0^{\infty} x^{\alpha-1}e^{-\frac{x}{\beta}}dx = \beta^{\alpha}\Gamma(\alpha), \tag{2.35}$$

and

$$\begin{aligned} E(X) &= \int_0^{\infty} x\left(\frac{x^{\alpha-1}e^{\frac{-x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}\right)dx \\ &= \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\int_0^{\infty}\frac{x^{\alpha}e^{\frac{-x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}dx \\ &= \frac{1}{\beta^{\alpha}\Gamma(\alpha)}[\beta^{a+1}\Gamma(\alpha+1)], \end{aligned} \tag{2.36}$$

and using exponential rules

$$E(X) = \alpha\beta. \tag{2.37}$$

The $\Gamma(1)$ vanishes because it is equal to 1 by direct integration. Now, for the second part we need to find the variance and to do that we need to recall that $V(X) = E[X^2] - [E(X)]^2$. It is clear now that the $[E(X)]^2$ is the key in order to finish the proof. So following the same steps

$$\begin{aligned} E(X^2) &= \int_0^{\infty} x^2\left(\frac{x^{\alpha-1}e^{\frac{-x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}\right)dx \\ &= \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\int_0^{\infty}\frac{x^{\alpha+1}e^{\frac{-x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)}dx, \\ &= \frac{1}{\beta^{\alpha}\Gamma(\alpha)}[\beta^{a+2}\Gamma(\alpha+2)] \\ &= \alpha(\alpha+1)\beta^2. \end{aligned} \tag{2.38}$$

The last step is to plug in the two findings in the variance's equation which gives us the following:

$$V(X) = \alpha(\alpha+1)\beta^2 - (\alpha\beta)^2 = ab^2. \tag{2.39}$$

□

Now, a new definition of the Chi-square distribution can be given.

Definition 2.4.2 (Chi-square Distribution)**.** If a random variable X follows a Γ distribution with parameters $\alpha = \frac{k}{2}$ and $\beta = 2$ then X is a Chi-squared distributed random variable with k *degrees of freedom* [8].

Fig. 2.4 illustrates the importance of k degrees of freedom. Different k leads the density function's curve to fluctuate.

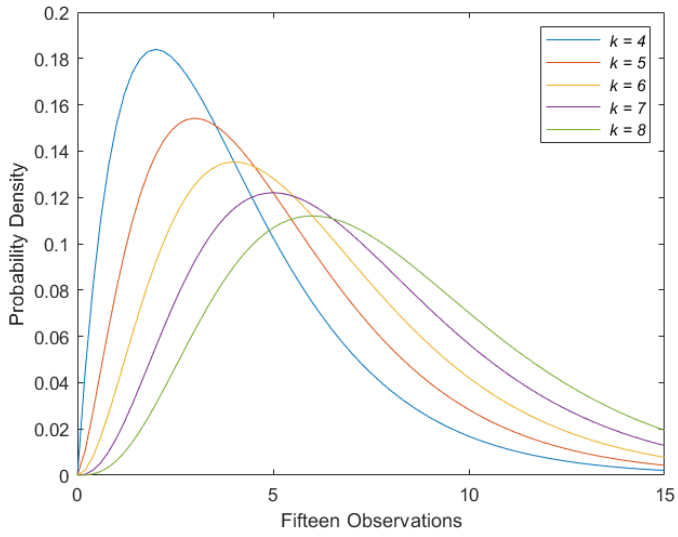

Figure 2.4: Chi-square density function with different k

2.5 Statistical tests

2.5.1 The Chi-square test of goodness-of-fit

The Chi-square test is used in order to examine either independence among two categorical variables or to show how good a sample fits to the distribution of a known population, in other words known as goodness-of-fit. Many tests as well as Chi-square test use the Chi-square distribution as the reference distribution in order to fit models [26]. Additionally, this thesis will later analyze that the reference distribution of the population follows or not the BL distribution. Since we have to do with a test, there is a null hypothesis H_0 which is that the observed or "true" distribution follows the BL distribution and an alternate hypothesis H_a which is the opposite of the null hypothesis. The most common test statistic formula for the Chi-square test of goodness-of-fit for the BL distribution is the following:

$$\chi^2 = n \sum_{d=1}^{9} \frac{(\frac{h_d}{n} - b_d)^2}{b_d}, \tag{2.40}$$

where n is the number of observations, h_d is the observed different frequencies for the digits 1 to 9, b_d is the BL distribution for each leading digit [6]. The test statistic follows a Chi-square distribution with 8 degrees of freedom under H_0. In addition, in case that $\chi^2 > \chi_{a,8}$, where a is the significance level, the H_0 will be rejected. The only disadvantage of this test seems to be the sensitivity that has when it comes to large sample size. As the the data of BL reject the H_0, the Chi-square test may have a problem to be a good goodness-of-fit test instrument [23].

The following example illustrates how BL can be used to exploit fraud. By using the values in the Table 2.2, someone can plug them into the corresponding parameters in eq. (2.40). That way, they will be able to find out if there is any data manipulation or not.

Example 2.5.1. *The leading digits from 1000 checks issued by seven companies were analyzed by an investigator. The observed frequencies corresponding to the leading digits 1, 2, 3, 4, 5, 6, 7, 8, 9 are 290, 180, 112, 95, 85, 69, 60, 53 and 56 respectively. If the observed frequencies are significantly different from the b_d, there is a possibility that the check amounts appear to result a fraud. Using a significance level of $a = 0.10$ and $k = 8$ to test for goodness-of-fit with Benford's Law, will the result suggest a possibility of fraud?*

First determine the H_0 and H_a:

- H_0: The observed distribution follows a BL distribution.
- H_a: At least one leading digit has a frequency that does not follow the BL distribution.

Table 2.2: Example's given data

Leading Digit	1	2	3	4	5	6	7	8	9
Observed Frequencies or h_d	290	180	112	95	85	69	60	53	56
Benford's Law: Distribution of Leading Digits or b_d	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%

The Table 2.2 is the key in order to solve the problem. Since there is enough information for the parameters in the eq. (2.40), we can calculate that $\chi^2 \approx 4.98$. In addition, the table of Appendix D in [27] can give the exact value of $\chi_{0.1,8} = 13.362$. We know that in order to reject H_0 the following must hold $\chi^2 > \chi_{a,8}$. However, this is not true since the inequality will be reversed in our case. Thus, there is not sufficient evidence to conclude that the checks suggest a fraud. In addition in the Fig. 2.5, it is clear how close the "true" distribution (orange bars) is to the Benford's Law distribution (green bars).

2.5.2 Distribution distance

An additional measure which we will use for testing the COVID-19 data sets uses as a base the Euclidean distance between the Benford's distribution and that of our data. This method can be seen as free from hypothesis testing and compatible with any sample size. Many studies have shown different approaches when using it, however, in our studies we chose the method from [15]. Thus, let $d = (\sum_{i=1}^{9}(b_d - \frac{h_d}{n})^2)^{\frac{1}{2}}$ be the Euclidean distance between the two sets. Then the modified test statistic d^* is as follows:

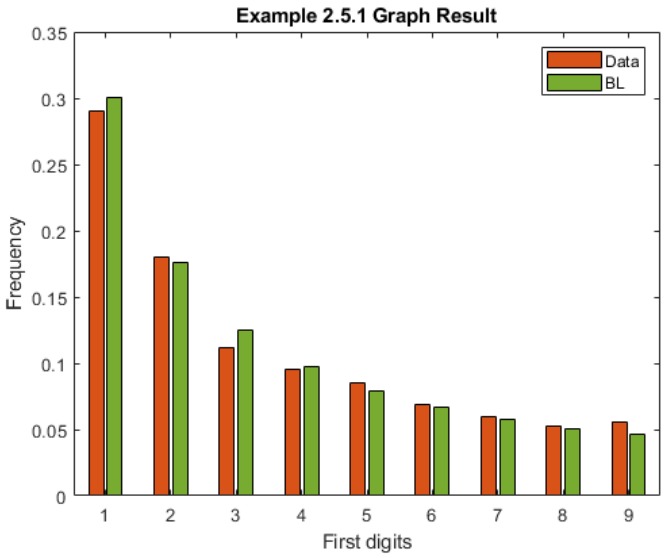

Figure 2.5: Example 2.5.1 visual comparison

$$d^* = \sqrt{n}\sqrt{\sum_{d=1}^{9}(b_d - \frac{h_d}{n})^2}, \tag{2.41}$$

where the variables have the same meaning as in the eq. (2.40) [6, 15]. Furthermore, the rejection regions regarding the significance levels can be found in the Table 2.3 [15].

Table 2.3: Rejection regions

Significance level	$a = .10$	$a = .05$	$a = .01$
Test statistic d^*	1.212	1.330	1.569

2.5.3 Mean Absolute Deviation

Mean Absolute Deviation (MAD) is a statistical test which ignores the number of records and is really useful for large size samples when statistical tests like Chi-square, are impractical for enormous real-world data. However, there is a problem when it comes to small samples. One of BL's problem is the false positives or in other words, data that are not biased do not follow BL distribution and vice versa. Consequently, this thesis will not use the standard model from [18] but an adjusted MAD model of [12] that is more effective with smaller data samples. In this study we use the MAD from [12], which is defined as follows:

$$MAD = \frac{1}{n}\sum_{1}^{n} f_i|x_i - \bar{x}|, \tag{2.42}$$

where n is the sample size, which equals to 9 for the first digits, x_i is the sample value, in our case this is the absolute value of the difference between the actual percentage and Benford's distribution of first digits, $\bar{x}$ is the mean or expected value, and f_i is the frequency, which is always 1 for this model. Additionally, the absolute symbol means that someone is interested only in the positive sign regardless if the deviation is positive or negative [4, 18, 1, 12].

There are not rejection regions as in the previous statistical tests, however, Nigrini presented some critical values for close conformity, acceptable conformity, marginally acceptable conformity, and nonconformity. The Table 2.4 presents the range of such critical values and the conclusions [18].

Table 2.4: Critical Values and conclusions for the MAD values

Digits	Range	Conclusion
First digits	0.000 to 0.006	Close conformity
	0.006 to 0.012	Acceptable conformity
	0.012 to 0.015	Marginally acceptable conformity
	Above 0.015	Nonconformity

Chapter 3

Methodology

3.1 The data

The COVID-19 quantitative secondary daily data that will be used in the methodology is exclusively obtained from the *Center for Systems Science and Engineering of John Hopkin's University* [5]. In our research we have decided to include four different countries to test their daily data reporting. These are Azerbaijan, Belarus, Greece, Serbia. It is worth mentioning that these countries have different ruling regimes. We took a close look at [10] for each of the countries to determine their democracy index. Thus, we can categorize Azerbaijan and Belarus as the countries with authoritarian regimes, and Greece and Serbia as countries with democratic regimes. The reason for choosing these countries that fall into these two categories is to see if there is a difference in the way that they report COVID-19 daily data to the public, with the help of BL. We will take a look at Sweden as well for contrast regarding data reporting as a democratic country. Furthermore, Table 3.1 illustrates the data sample periods that are going to be examined for each country daily.

Table 3.1: Data sample periods

	First category		Second category		Third category	
Countries	Start	End	Start	End	Start	End
Azerbaijan †	Mar 1, 2020	Mar 14, 2021	Mar 1, 2020	Mar 28, 2020	Mar 29, 2020	Mar 14, 2021
Belarus ††	Feb 28, 2020	Mar 14, 2021	-	-	-	-
Greece $*$	Feb 26, 2020	Mar 14, 2021	Feb 26, 2020	Mar 22, 2020	Mar 23, 2020	Mar 14, 2021
Serbia $**$	Mar 6, 2020	Mar 14, 2021	Mar 6, 2020	Mar 17, 2020	Mar 18, 2020	Mar 14, 2021

We decided to split the time frame into three categories according to [6] because it will be easier for us to justify certain concepts. Firstly, the pandemic follows exponential growth when it comes to the first reported cases, meaning that the Chi-square test should not reject the null hypothesis. That way we are able to immediately show some evidence of false or

†Source: https://nk.gov.az/en/article/747/

††There is no source that indicates lockdowns.

$*$Source: https://primeminister.gr/en/2020/03/22/23619

$**$Source: https://www.srbija.gov.rs/vest/en/151641/

non-false reporting. Secondly, after government involvement and new enforced restrictions in each of our countries, we expect that here the data will have the most disturbance between the true distribution and the BL distribution. Last but not least, we will look at the whole reported data, where we expect it to act in the same way as the period with restrictions, since the period sample before that is not that big. Thus, our categories are as follows:

- First category: the full time sample which consists of all the data, starting from the first recorded case in each country.
- Second category: the period that takes into consideration the data until the first government intervention that affected every citizen (i.e. curfew, national lockdown, mandatory rules) which were implemented against the spread of COVID-19 in each country.
- Third category: the period taking the data after the first government intervention against the spread of the virus.

We decide to analyze the daily data from all countries until 14th of March 2021. The time period of the three categories were justified by using multiple website sources, listed in the footnotes of the previous page. It is worth mentioning that we have not found any source that would show that Belarus has implemented any restriction to reduce the spread of the virus, thus we omit Belarus from second and third category analysis in this thesis. In addition, daily data that has 0 recorded new daily cases are omitted from our analysis.

3.2 Testing thesis method

Before we proceed with our data analysis, we will look at the data that were given in [6] and conduct our first tests on them. In this manner, we will test if our methodology conforms to theirs by evaluating whether our results are in accordance with their findings.

We found through our testing that the results from [6] for China and the US differ from ours. Although the sources that we used and the time intervals are the same, we can't justify the difference between the data that the authors gave in their study. Additionally, it is worth mentioning that the addition of the pre-lockdown and post-lockdown periods should sum up to the full sample. It is quite clear that these two periods that are given in [6] regarding China, do not sum up to 705, but to 733. There is not any justification if this is intended or not in the corresponding article. From the above mentioned comments, we can only assume that the data sources used might have been updated, since their and our extraction periods differ from each other.

When it comes to Italy, we validate the data with theirs by using the bulletins from *Dipartimento della Protezione Civile* [7]. Thus, our results are in accordance with theirs. This indicates that our method is accurate enough, but does raise questions about the data used for the other two countries.

Consequently, in order to compare our findings to theirs, we need to define the following variables that represent our findings: n_m for sample size, χ^2_m for the Chi-square and d^*_m for the distribution distance. Table 3.2 represents the dissimilarities between our and the authors'

findings. These, however, cannot be fully explained. As we have mentioned above, one of the reason for the differences could be due to updates in certain data cells after the extraction date presented in [6] or possible differences in the methodology that was not fully described in their paper.

Table 3.2: Comparison of results with those from [6]

Countries	Time	n, [6]	χ^2, [6]	d^*, [6]	n_m	χ^2_m	d^*_m
China	Full Sample	705	25.334	1.718	733	22.874	1.7397
China	Pre-lockdown period	581	16.036	1.166	644	15.732	1.4178
China	Post-lockdown period	145	23.785	1.891	89	9.6484	1.3111
Italy	Full Sample	980	18.129	1.689	980	18.129	1.6899
Italy	Pre-lockdown period	359	4.9964	0.65	359	4.9964	0.6503
Italy	Post-lockdown period	621	39.613	2.312	621	39.613	2.3124
U.S	Full Sample	5479	15.19	1.074	5541	16.739	1.1343
U.S	Pre-lockdown period	1867	11.395	1.314	1803	5.0095	0.8049
U.S	Post-lockdown period	3612	20.029	1.246	3738	18.745	1.2568

In addition, we conduct another test for the MAD by testing our method to the data's time period given in [1]. Given the same source as theirs, we conduct the test only for three of the countries used in their paper, which are Albania, Belgium, and Turkey. There is no specific reason why we chose those specific countries. Furthermore, we conclude that our method gives a slightly numerical difference for Albania and Belgium but a higher numerical difference when it comes to Turkey. There is not an absolute explanation for getting different values, but this could be due to updates to certain data cells after their extraction date or possible differences in the methodology. Furthermore, Table 4.2 indicates our results and their dissimilarities compare to the authors' results.

Table 3.3: Comparison of obtained MAD results

Countries	MAD, [1]	MAD
Albania	0.035	0.041
Belgium	0.019	0.0165
Turkey	0.067	0.045

3.3 Data Analysis

The data analysis carries through with the three chosen statistical tests, which are the Chi-square test goodness-of-fit, the distance distribution test, and MAD. In this way we analyze if the corresponding data follows a BL distribution or not, and to what extent. It is worth mentioning that our hypothesis are:

- H_0: The observed distribution follows a BL distribution.
- H_a: At least one leading digit has a frequency that does not follow the BL distribution.

We have $k = 8$ df and a significance level of $a = 0.1$ for the Chi-square test. This results in the rejection of the null hypothesis if $\chi^2 > \chi^2_{0.1,8} = 13.3616$ [27]. We refer to Table 2.3 for the d^*'s rejection regions and to Table 2.4 for the valuation of MAD results. In addition, we use certain software like MatLab, R and Excel for calculations, graphing and testing. For the calculations of the χ^2 we use the benford.analysis package [2].

Chapter 4

Results

As we have tested the data by using R and Excel, we obtained the results presented in Table 4.1 along with the Fig. 4.1. What was observed immediately are the odd results regarding the countries with authoritarian regimes, which are Azerbaijan and Belarus. Both of them reject the null hypothesis of Chi-square test. Oddly, the Chi-square does not reject the null hypothesis for Azerbaijan in the First Category data, but this might attribute to the very low amount of data, which is also observed for other countries in the first category. In addition, the d^* test has much larger values than for the other two democratic countries, Greece and Serbia, and which exceeds any of the rejection regions presented in 2.5.2. The MAD results show nonconformity for the BL for Azerbaijan in the Second and Third Category analysis, however it shows marginally acceptable conformity for the First Category analysis. Belarus on the other hand does not conform with the BL according to MAD and the Chi-square goodness-of-fit test. Moreover, by looking at Fig. 4.1, it is obvious that Belarus does not follow the BL distribution.

In addition, the countries with democratic regimes, which are Greece and Serbia, both reject the null hypothesis of the Chi-square test as well, but do perform better than the authoritarian countries when it comes to the d^* test. We found that for both countries the d^* shows that they conform to the BL through the Second Category data analysis, while First and Third Category data are in the rejection region. When looking at the MAD results, the performance is much better for both countries in the First Category data analysis, however, Greece shows conformity in both First and Third Category, while Serbia on the other hand does not conform in any of the categories. In contrast with the authoritarian countries, Fig. 4.1 shows that the "true" distribution of the data is really close to the BL distribution for both Greece and Serbia in the First Category data.

The last democratic country that was considered in this analysis is Sweden. In addition, there is only the First Category for Sweden since there was no lockdown period imposed. According to Table 4.1, there is evidence that all of the tests reject the null hypothesis when it comes to the First Category.

Table 4.1: Chi-square test, data acquired from [5]

Countries	Time	n	χ^2	d^*	MAD
Azerbaijan	First category	367	36.986	2.097	0.015
Azerbaijan	Second category	16	1.2119	0.412	0.018
Azerbaijan	Third category	351	38.031	2.119	0.016
Belarus	First category	359	202.14	5.316	0.043
Greece	First category	375	21.888	1.782	0.014
Greece	Second category	25	16.297	0.80	0.022
Greece	Third category	350	23.946	1.759	0.014
Serbia	First category	366	22.763	1.842	0.017
Serbia	Second category	9	15.927	1.09	0.057
Serbia	Third category	357	22.221	1.904	0.019
Sweden	First Category	293	44.78	2.396	0.022

Given the results that we have obtained for Sweden, we decided to inspect the data further, since we are more familiar with the governmental sources of Sweden. Thus, we have done the same analysis, but with the data published by *Folkhälsomyndigheten (Public Health Agency of Sweden)* [11], which in result gives us more data, with detailed information of confirmed cases through all of the 21 counties in Sweden. What was observed is that MAD showed acceptable conformity, the best that we have observed during our testings (see Table 4.2). The d^* and χ^2 however are still high, but not larger than that of Belarus from Table 4.1. The reason of the high value of the χ^2-test is explained by the large sample that we used, as this test is very sensitive when it comes to large data samples. When looking at Fig. 4.2, the data conforms closely with the BL, to much higher extent than in Fig. 4.1. Conclusions based on these obtained results, as well as some ideas for future investigations, follow in Chapter 5.

Table 4.2: Chi-square test, data acquired from [11]

Countries	Time	n	χ^2	d^*	MAD
Sweden	First Category	7265	83.075	4.059	0.008

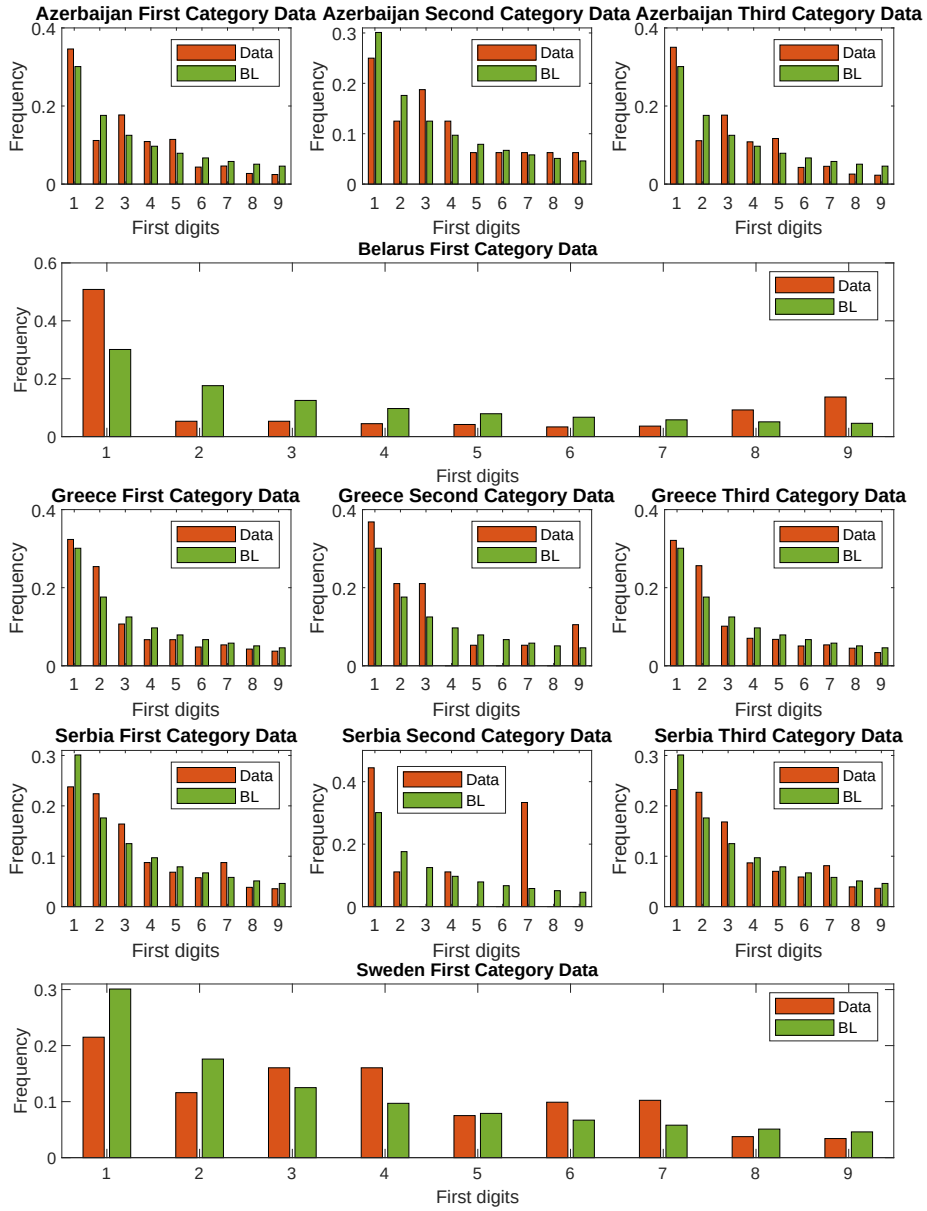

Figure 4.1: Graph results comparing our data with BL for each country, data acquired from [5]

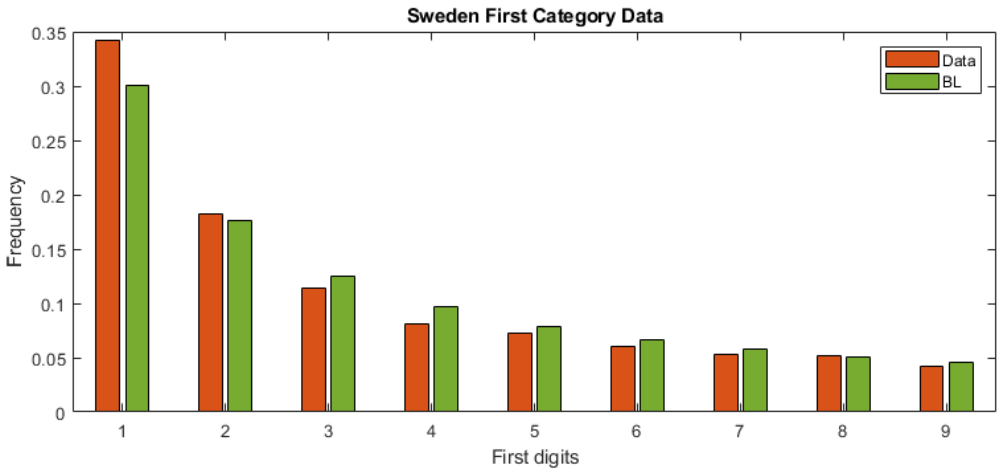

Figure 4.2: Results for Sweden for data acquired from [11]

Chapter 5

Conclusion

5.1 Thesis summary

In this thesis, we focus on the trustworthiness of the reporting of COVID-19 daily cases in Azerbaijan, Belarus, Greece, Serbia and Sweden. The BL is the main tool of this investigation which is well known for its fraud detection properties [18] when it is used in the context in other statistical tests. As a result, we use the three following statistical tests: Chi-square goodness-of-fit, distribution distance and MAD. The aim of this study is to provide information to the public about any inconsistency that might have occured between the "true distribution" and the BL distribution for the data of the corresponding five countries.

In the first chapter, we gave an introduction about BL's historical background and a reference about the authors who were behind this important law. Furthermore, the second chapter is about the BL's generalization, proof and limitations. It was essential to write some of the distributions before explaining the Chi-square goodness-of-fit test for this study's sake. Further, the chosen three statistical tests of this thesis were conducted.

In Chapter 3, we introduced three different categories which correspond to three different time intervals. We used different trustworthy sources, e.g. official government bodies, in order to be as precise as possible with the starting and ending periods of each category. After this, we conducted a test to check if our methodology was compatible with the methodology that was used in [6]. However, when it comes to the results only Italy's were identical to ours. The results regarding China and the U.S were different. Consequently, this difference may be due to potentially updates in the data cells of COVID-19 daily cases of [5] after the authors' extraction date. Moreover, we applied our methodology to [1] for the MAD test. Consequently, our results were appreciably similar to their results for Albania and Belgium, but not for Turkey. Last but not least, Chapter 3 ends with the data analysis which consists of details about the hypotheses, the rejection regions of the statistical tests, and the software used.

When it comes to the Results, we splat the five countries into two categories: countries with democratic regimes (Greece, Serbia, Sweden) and countries with presidential republic regimes (Azerbaijan, Belarus). Our data were taken from the *Center for Systems Science and Engineering of John Hopkin's University* [5]. We concluded that the results for Azerbaijan were slightly concerning and may need further investigation since certain categories rejected the

null hypothesis. When it comes to Belarus, all of the tests rejected the null hypothesis, which makes Belarus a potential case of COVID-19 data misinformation. Additionally, Belarus's χ^2 value was enormous which is odd since for the other three countries that almost had an identical sample size, the value of χ^2 was not as extreme. Greece and Serbia gave problematic results for certain categories but they are surely closer to the BL distribution than the authoritarian countries, according to Fig. 4.1. However, further investigation may be required. Lastly, Sweden rejected the null hypothesis for every possible statistical test. This made us realize that there could be a problem with the given data for Sweden from the *Center for Systems Science and Engineering of John Hopkin's University* [5] since the sample size is small. Additional test was done using the data taken from *Folkhälsomyndigheten (Public Health Agency of Sweden)*[11] and gave results showing distribution almost identical to the BL one but with far more observations (see Chapter 4).

5.2 Future work

Even though almost all of the countries rejected our tests for their corresponding categories, we believe that Greece and Serbia were the countries with the best fitting to the BL. This conclusion is based on graphically presents results (see Chapter 4), but also based on the statistical test results of which many were close or inside the value intervals of the rejection region. However, these two countries could act in a more sufficient way if there was a bigger sample size. We cannot say the same for Azerbaijan since the results were far away from the rejection regions when it comes to the first and the third category. In addition, in our opinion, Belarus needs some further investigation since most of the tests failed and there is evidence which indicates potential misinformation to related to COVID-19 data provided to the public. Thus, for future research, we propose several ideas that were not tested in this thesis:

- As it is clear that the enormous sample size of Sweden given by the *Folkhälsomyndigheten* fits almost perfectly to the BL distribution, it will be better to test this paper's methodology in a larger sample size with far more observations.
- Someone can try to test more countries with presidential republic regimes and see if they follow a similar pattern, as the two countries in this research.
- Lastly, there are more statistical tests out there that will prove to be crucial, to both be applied and tested in a similar research. For example, the Kolmogorov–Smirnov test, the Z test or similar.

Chapter 6

Reflection of objectives in the thesis

A summary of the objectives achieved in this study will be presented in this chapter.

6.1 Objective 1: Knowledge and understanding

This study indicates that new knowledge was attained. It shows how applied mathematics work in this sort of research. Some related concepts that helped to conduct this analysis, were in the context of calculus and statistics. In Chapter 2, related concepts to calculus were studied in order to prove BL, e.g. Laplace transform. In addition, in Chapter 2 and 3 statistics are used in order to describe some distributions and to give an insight into the statistical tests used in the research. Moreover, in Chapter 4 we showed and discussed our results. This has shown that we gained a deep understanding of BL as well as of the three statistical tests used. In Section 5.2, we have proposed few future directions which has further showed our understanding of the material covered in this thesis. Computer skills like programming language knowledge, LaTeX, Excel, etc. were demanded for a smooth outcome of the thesis.

6.2 Objective 2: Methodological knowledge

We presented the subject after giving some information about the theoretical background first. A methodological knowledge is demonstrated by using various reliable references, tables, figures and examples in order to make the reader feel more comfortable with the concepts of the thesis. This was achieved by using MATLAB for graphing and R and Excel for the calculations.

6.3 Objective 3: Critically and Systematically Integrate Knowledge

Information is taken from many different sources. Starting references provided by our supervisor were built upon which ended up exploring many scientific articles and books mainly referred to Benford's Law for being able to further explain this specific concept.

6.4 Objective 4: Independently and Creatively Identify and Carry out Advanced Tasks

The very first step of the thesis was to find a research question but also to do a research on the corresponding topic. We came to an agreement to include several important chapters: *Introduction* which gives details about the historical background, purpose, aim and methodology. Followed by the *Theoretical Consideration*, which includes information about BL, statistical distributions and statistical tests. *Methodology* where we compared our methods to other authors' papers, and lastly *Results* and *Conclusion* where we presented our methodology's results for five countries and made correct conclusions based on these results. The guidance provided by the supervisor helped tackling certain difficult parts of the thesis and led to a well-structured project report.

6.5 Objective 5: Present and Discuss Conclusions and Knowledge

The thesis is not hard to be followed by people that are neither having a mathematical background, nor a statistics background. For some of the concepts, the reader might need to do some individual reading to attain a deeper understanding by using our sources in the Bibliography section, but the general idea is easy to be followed in our opinion. A lot of sources are used in order to explain the concepts in details as much as possible. Figures, tables and some short numerical examples can be found in the thesis. Consequently, the reader will be able to understand most of the topics even without having a deep knowledge of the concepts that are used. An oral presentation of the work that has be done will take place in June 2021 when everyone is welcome to attend and ask questions regarding the concepts and results that can be found in the thesis. Lastly, noteworthy is that we have been practising on how to present our results both orally (discussing during meetings) and written (sending a draft before the meeting) for each meeting with the supervisor.

6.6 Objective 6: Scientific, Social and Ethical Aspects

All of the sources used in the thesis are properly cited in the study. In addition, the data and R package used can be found in the Bibliography. When it comes to ethics, the work is done with caution and avoiding direct accusations. Lastly, everyone that helped us to achieve our goals and to complete this thesis will be mentioned in the Acknowledgements.

Bibliography

[1] A.Kilani, G.P.Georgiou, Countries with potential data misreport based on Benford's law. *J Public Health (Oxf)*, 2021, https://doi.org/10.1093/pubmed/fdab001.

[2] C.Cinelli, (2015, November 22), benford.analysis, https://carloscinelli.com/software.html.

[3] C.Durtschi, W.Hillison, C.Pacini, The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data, *Journal of Forensic Accounting*, R.T Edwards, Volume **1524-5586**, 2004, Pages 17-34, https://www.researchgate.net/publication/241401706_The_Effective_Use_of_Benford's_Law_to_Assist_in_Detecting_Fraud_in_Accounting_Data.

[4] C.S.Azevedo, R.F.Goncalves, V.L.Gava, M.M.Spinola, A Benford's Law based methodology for fraud detection in social welfare programs: Bolsa Familia analysis, *Physica A: Statistical Mechanics and its Applications*, Volume **567**, 2021, Pages 1-13, ISSN 0378-4371, https://doi.org/10.1016/j.physa.2020.125626.

[5] CSSE John Hopkins University, Time Series COVID-19, n.d., Retrieved 15 March 2021 from https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series.

[6] C.Koch, K.Okamura. Benford's Law and COVID-19 reporting, *Economics letters*, Volume **196**, 2020, Pages 1-4, https://doi.org/10.1016/j.econlet.2020.109573.

[7] Dipartimento della Protezione Civile, dati-regioni, n.d., Retrieved 13 April 2021 from https://github.com/pcm-dpc/COVID-19/tree/master/dati-regioni.

[8] D.Wackerly, W.Mendenhall, R.Scheaffer. *Mathematical Statistics With Applications*, Thomson Learning Emea., 2007, ISBN 978-049-53-8508-0.

[9] E.Costas, V.Lopez-Rodas, J.F.Toro, A.Flores-Moya. *Aquatic Botany*, Volume **89**, 2008, Pages 341-343, https://doi.org/10.1016/j.aquabot.2008.03.011.

[10] Economist Inteligence Unit (2020). *Democracy Index 2020: In sickness and in health?*, 2020.

[11] Folkhälsomyndigheten, Bekräftade fall i Sverige – daglig uppdatering, n.d., Retreived 30 April 2020 from https://www.folkhalsomyndigheten.se/smittskydd-beredskap/utbrott/aktuella-utbrott/covid-19/statistik-och-analyser/bekraftade-fall-i-sverige/.

[12] G.G.Johnson, J.Weggenmann, Exploratory research applying benford's law to selected balances in the financial statements of state governments, *Academy of Accounting and Financial Studies Journal*, Volume **17**, 2013, Pages 31-44.

[13] J.Adrien. Benford's Law, Master's thesis, Imperial College of London, 2001.

[14] J.Goldbeck. Benford's Law Applies To Online Social Networks, 2015, https://10.1371/journal.pone.0135169.

[15] J.Morrow, Benford's Law, Families of Distributions and a Test Basis, Centre for Economic Performance, LSE, 2014, https://ideas.repec.org/p/cep/cepdps/dp1291.html.

[16] M.Ausloos, C.Herteliu, B.Ileanu. Breakdown of Benford's law for birth data, *Physica A: Statistical Mechanics and its Applications*, Volume **419**, 2015, Pages 736-745, ISSN 0378-4371.

[17] M.Cong, B.Ma. A Proof of First Digit Law from Laplace Transform, 2019, https://doi.org/10.1088/0256-307X/36/7/070201.

[18] M.Nigrini, Benford's Law: Applications for Forensic Accounting, Auditing and Fraud Detection, John Wiley and Sons, 2012 ISBN 978-111-81-5285-0.

[19] P.Manoochehrian, F.Rachidi, W.Schulz, M.Rubinstein, G.Diendorfer. *Benford's law and lightning data*, 2010.

[20] R.Kissell, J.Poserina. *Optimal Sports Math, Statistics, and Fantasy*, Academic Press, 2017, ISBN 978-012-80-5163-4, https://www.sciencedirect.com/science/article/pii/B978012805163400013X.

[21] S.J.Miller. *Benford's Law: Theory and Applications*, Princeton University Press, 2015, ISBN 978-069-11-4761-1.

[22] S.Newcomb. Note on the Frequency of Use of the Different Digits in Natural Numbers, *American Journal of Mathematics*, The Johns Hopkins University Press, Volume **4**, 1881, Pages 39-40, ISSN 0002-9327, https://www.jstor.org/stable/2369148.

[23] S.C.Y.Wong, *Testing Benford's Law with the First Two Significant Digits*, University of Victoria, 2010.

[24] T.P.Hill. A Statistical Derivation of the Significant-Digit Law, *Statistical Science*, Volume **10(4)**, 1995, Pages 354–363. https://doi.org/10.1214/ss/1177009869.

[25] T.P.Hill, Base-Invariance Implies Benford's Law. *Proceedings of the American Mathematical Society*, Volume **123(3)**, 1995, Pages 887-895 https://doi.org/10.2307/2160815.

[26] T.M.Franke, C.A.Christie, T.Ho, The Chi-Square Test Often Used and More Often Misinterpreted, *American Journal of Evaluation*, Volume **33**, 2012, Pages 448-458, https://doi.org/10.1177/1098214011426594.

[27] T.L.Vanpool, R.D.Leonard, *Quantitative Analysis in Archaeology*, John Wiley & Sons, 2011, ISBN 978-140-51-8951-4.