Universidad de Valladolid
Science Faculty

Final degree Project

Degree in Statistics

Benford's Law. History, mathematical justification and applications.

Author: Naila Talavera Palacios
Tutor: Agustín Mayo

Table of contents

Abstract

Newcomb-Benford's Law, also known as "the first significant -digit Law " or "the Law of Anomalous Numbers ", is based on an observation from which a law was formalized about the distribution of the first significant digits in positive numerical data, where the probabilities of the most significant figures are not uniformly distributed, as might be expected. In addition, it receives these other names because it shows how the lower digits occur in nature more-often than the higher ones, indicating the absence of this event a possible risk of abnormal duplications and anomalies in certain datasets, and this is why it is a key tool in fields like fraud detection.

In this sense, the issue of this project is to summarize bibliography with the purpose of showing the mathematical, statistical and empirical frameworks concerning this Law. We will study the distribution of the first significant digits, point out how to apply the formulas and we will interpret and check its effectiveness with the results obtained using two different datasets. Finally, we will also review the applications of this law in our day-to-day.

Keywords: Newcomb-Benford's Law; First or significant -digit Law; The Law of Anomalous Number; distribution of the first significant digits; fraud detection.

Resumen

La ley de Newcomb-Benford , también llamada "la ley del primer dígito" o "ley de los números anómalos", está basada en una observación a partir de la cual se formalizó una ley sobre la distribución de los primeros dígitos significativos en conjuntos de datos numéricos positivos, donde las probabilidades de las cifras más significativas no están distribuidas de manera uniforme, como cabría esperar. Además, recibe estos otros nombres ya que muestra cómo los primeros dígitos ocurren en la naturaleza con mayor frecuencia que los últimos, indicando la ausencia de este suceso en ciertos conjuntos de datos un alto riesgo de que este contenga anomalías, y es por eso que es una herramienta clave en campos tales como la detección de fraude.

En este sentido, este proyecto se basará en la recopilación de bibliografía acerca de la ley de Newcomb-Benford con el objeto de mostrar su marco matemático, estadístico y empírico. Estudiaremos la distribución de los primeros dígitos significativos y lo aplicaremos a dos conjuntos de datos reales . Por último, revisaremos también las aplicaciones de dicha ley en nuestro día a día.

Palabras clave: Ley de Newcomb-Benford, ley del primer dígito, ley de los números anómalos, distribución de los primeros dígitos significativos, detección de fraude.

1. Introduction

Benford's Law gives us the opportunity to wonder whether there are patterns to the numbers that we use daily and these patterns could help us identify whether a dataset has been manipulated or not.

This Law consists of the description of the distribution of the first significant digits, which ensures that the initial digits of the numbers are not equiprobable in datasets which conform to the Law. Therefore, the probability of occurrence of digits, in decimal notation, is not uniform. Those numbers starting with the digit one occur more-often than those starting with two and so on, up to 9, which is the least frequency. This indicates that as the value of the first digit increases, the more unlikely it is. This fact, which is accepted as a law, can be applied to large datasets made up of positive numbers and related to the natural world or social data.

The Law was named after Benford, but he was not the first to study it. Therefore, we prefer to refer to this law as the Newcomb-Benford's Law, as other authors did, whose acronym will be **NBL** for short, in order not to forget Newcomb's merit, who was the first to publish a paper related to the Law. Moreover, NBL was forgotten after its discovery, maybe because it was non-intuitive and not well recognized because of the absence of mathematical foundation. However, some decades ago it regained its interest due to the number of its applications increased considerably.

Undoubtedly, the most important application of the Law is its use as a possible mean to assist in the detection of erroneous or fraudulent data as we have mentioned. If the law is not verified this may be due to the presence of inaccurate or retouched data. Evidently, the non-conformity to NBL in a given dataset is not a sufficient proof of the existence of these irregularities. Nevertheless, it constitutes a good indication to justify a further detailed inspection.

Nowadays, NBL has widespread use in other areas such as forensic accounting, auditing, etc. In many countries, especially in Venezuela and Mexico, it has been used as the main tool in the detection of electoral fraud. The treasury of the USA also uses it to detect fiscal fraud, auditors to analyse accounting statements…

NBL also makes predictions about the distribution of second digits, third digits, digit combinations, and so on. In fact, NBL describes the probability distribution of all digits. However, it is not universally applicable, we need data from any distribution taking values over a wide range, covering different orders of magnitude and being reasonably smooth, then it will tend to be Benford. For this reason, it has received many criticisms, as it cannot be used indiscriminately. Its mathematical background has already been proved and the empirical evidence is obvious, its practical potentialities have also been recognized, and that meanwhile this empirically derived Law can be considered theoretically well-analyzed, but we need a connection between both backgrounds, so let us summarize it.

2. Biography/ History

In 1881, the mathematician and astronomer Simon Newcomb [25] wrote a paper titled "*Note on the frequency of use of the different digits in natural numbers"* in *the "Americal Journal of Mathematics",* being the first to publish a paper related to "The first significant digit Law". He noticed that the nine digits, excluding zero, do not occur with equal frequency looking at the pages of the logarithms' book because the first pages were more worn than the last ones due to their use, so from Newcomb's personal observation a new Law was going to be stated.

He concluded his study using statements like "natural numbers are to be considered as quantities' ratios", " *The law of probability of the occurrence of numbers is such that all the mantissae of their logarithms are equally likely*", and he proposed the following formula:

$P(N) = \log(N + 1) - \log(N)$, with P(N): the probability of a single number *N as* the first digit.

but he did not investigate it further. Newcomb did not include the most representative formula in his paper which we will see later, although he used it in its study.

Frank Benford [3], who was a physicist from Pennsylvania, took over Newcomb's work and extended it, but his research only dealt with his hobby, which was mathematics. In 1938, he published the paper titled "The Law of Anomalous Numbers paper" .

The first step taken by Benford was to perform an analysis of the first digits of the positive numbers in 20 data tables by hand using a total of 20,229 records. He selected data from several datasets such as the areas of rivers, population figures or atomic weights of elements, and about one-half of his datasets did not exhibit the property. His work illustrates the empirical observation about the occurrence of each of the digits showing that smaller digits occur more often than greater ones as the first digits in those datasets. This concept is contrary to the common intuition as if they were distributed uniformly.

In 1972, the economist Hal Varian [31] suggested NBL could be used to prove if data are fraudulent and in that case, it will not conform to NBL. Some years later, in 1995, its mathematical foundations were formalized thanks to T. P. Hill [14, 15, 16], where its main properties were established.

3. Mathematical justification.

We are going to explain now the mathematical framework thanks to the contributions mainly from T.P. Hill [14, 15, 16], Arno Berger [4] and Allaart [2] among others, since their studies on NBL have turned out to be vital for its correct and complete explanation. However, as a correct introduction, let us explain first the proper meanings of **significant, significand** and **mantissa** due to these words are going to appear throughout the explanation:

Definition 3.1. The first **significant** digit of a real number $x > 0$ is the first digit different from zero that appears in decimal expansion, e.g., the first significant digit of 301 and 0.301 is 3 in both cases. This definition is applicable to other digits.

In scientific notation, the **significand** of a real number is its coefficient when we express it in floating point, e.g., 20= 2*10 so its significand is 2.

Definition 3.2. The Significand S(x) of $x> 0 \in \mathbb{R}$ is the unique real number such that, assuming base 10:

$$x = S(x)10^k \quad S(x) \in [1, 10) \quad \exists k \in \mathbb{Z}$$

The order of magnitude, k, represents the number that more varies according to its absolute value.
Remember that it is related to the first-digit, if we go now to the first-two digits the interval's definition would change to $S(x) \in [10, 100)$.

In American English, what we have defined as significand is denoted as mantissa, Currently, exists an open debate on the merging of the meanings of significand and mantissa, but the "traditional" meaning of mantissa is:

Definition 3.3. The mantissa is the difference between the logarithm of a number and its integer part, log(x) – [log(x)], where if we represent the mantissa of the logarithm base 10 of a number x by m(x), it has the following property: m(x)= m(10x). That is, the mantissae are cyclic, circular data and can take values on the unit circle centred at (0,0) [1].

The relationship between mantissa and significand lies in m(x)= log(S(x)).

All references in this paper to logarithms will be to the base 10, if not otherwise indicated.

3.1 The first digit distribution

Taking into account the above definitions, we could see that the significant digit of x is the same as the significant digit of S(x). However, we only need to evaluate one interval for S(x), [d, d+1) and for x there are several: [d, d+1), [10d, 10(d+1)), [10^2d, 10^2(d+1)), etc, where d denotes the different digits. Thus, we choose the only interval for S(x) because we are interested in the first digit and it is common for both, as:

$$log(x) = log(S(x)10^k) = log(S(x)) + log(10^k) = log(S(x)) + k$$

The probability of the first digit d, using the uniform distribution as the natural way to think about digits, is the probability of the significand S(x) ∈ [d, d+1), d ≤ S(x) < < d+1, and on a logarithmic scale, log (d) ≤ log(S(x)) < log (d+1), where we obtain log(d+1) – log(d) = lg(1+d^{-1}) , which is the interval width [11].

Definition 3.1.1. The distribution of the first significant digit.

$$P(D_1 = d) = log(1 + d^{-1}) \quad \forall d = 1, 2, 3, \ldots, 9$$

where D1 is the first significant decimal digit.

These significant digits are dependent as they follow logarithmic laws and their results show that 30.1% of the numbers have as first digit 1. The first digit 2 occurs 17.6 percent of the time, and if we sum both results we obtain a percentage close to fifty, 47.12%, so almost half of the observed numbers which would conform perfectly to NBL will have in its first place the number one and two. This logarithmic pattern continued up to 9, where only 4.6% of the numbers would have the first digit 9. They are obviously not uniformly distributed as we can see in the graphic below.

We use the logarithmic scale since it allows us to work with a wide variety of quantities, as a condition to achieve a good fit to the Law, because the logarithm reduces them to a more manageable range.

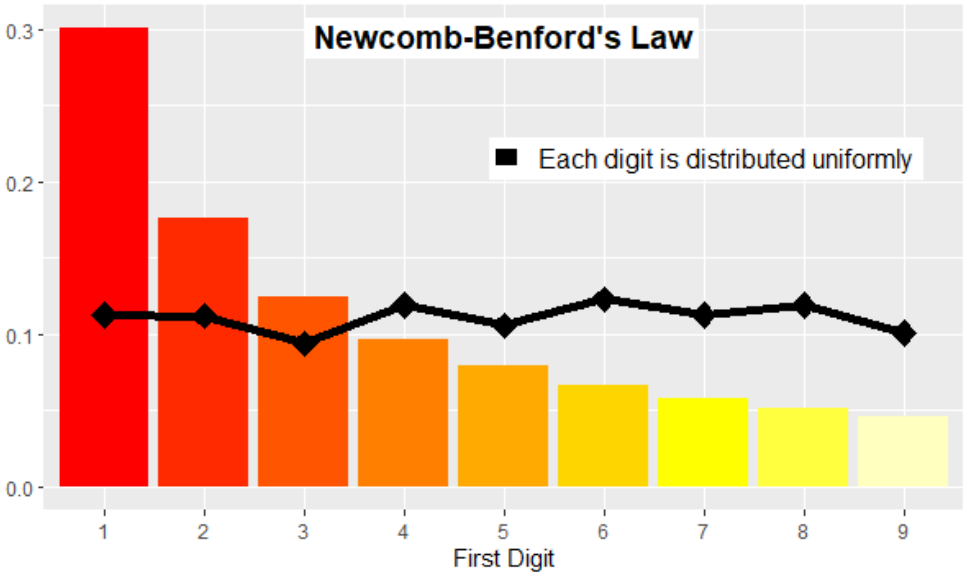

Figure 3.1. *Benford's Law plot using heat.colors in R. The higher the probability of the first significant digit, the redder the bar. The black line shows the case if the numbers were uniformly distributed, which has been calculated from a random sample of a U(0,1) with n=1000.*

Definition 3.1.2. Its distribution function is :

$$P(D \leq d) = \sum_{1 \leq d_* \leq d} P(D = d_*) = \sum_{1 \leq d_* \leq d} log(1 + d_*^{-1}) =$$

$$= log(\prod_{1 \leq d_* \leq d} (1 + d_*^{-1})) = log(2 \times \frac{3}{2} \times \ldots \times \frac{d+1}{d}) = log(d+1)$$

with P(D≤ 9)=1 [20].

Example 3.1.3. The number x ∈ [0, 10) begins with digit 1 if 1 ≤ x < 2 , x begins with digit 9 if 9 ≤ x < 10. Applying logarithms, log (1) ≤ log (x) < log (2) and log (9)≤ ≤ log (x) < log (10). The first interval is wider than the second interval, although before applying logarithms they had the same length, but this hinges on the idea that the first digits are not evenly distributed, thus as log(x) is distributed uniformly, it is more likely to start with 1 than with 9 because its interval width is larger.

In NBL, its mantissae are uniformly distributed between 0 and 1; from this, we can now understand what we mentioned in Section 2, *'The Law of probability of the occurrence of numbers is such that all mantissae of their logarithms are equally probable''*, as Newcomb stated.

Following with this idea, now suppose a random variable X which represents the number whose logarithm is searched for [20], then:

Property 3.1.4. *X satisfies NBL ⇔ log X – [log X] ~ U(0, 1),* where [.] is the floor function.

In order to finish this section, let us show now NBL's predictions about the distribution of second digits and first-two digits ,as this law describes the probability distribution of all digits and we are going to mention them later:

Definition 3.1.5. Probability of the second-place digit:

$$P(D_2 = d_2) = \sum_{d_1=1}^{9} log(1 + (10d_1 + d_2)^{-1}) \quad \forall d_2 = 0, \ldots, 9$$

Definition 3.1.6. Probability of the first-two digits:

$$P(D_1 D_2 = d_1 d_2) = log(1 + \frac{1}{d_1 d_2}) \quad d_1 d_2 \in \{10, \ldots, 99\}$$

3.2. The Generalized Significant Digit Law

Benford claimed that "*mere man counts arithmetically,{1,2,3,..} and it is not the natural number scale, nature counts* e^0 , e^x , e^{2x} *, and so on*". Since it appears that many natural functions are of the logarithmic form in base e, neither Newcomb nor Benford provided any theoretical basis. In 1995, T.P. Hill [16] provided a generalization of the first significant digit law stated in terms of the significant digit functions $\{D_b^{(i)}\}$, which for $b \in \mathbb{N} \setminus \{1\}$, $D_b^{(i)}(x)$ is the i^{th} significant digit of x when represented base b.

Definition 3.2.1. For the general case of any base, ∀ integer b>1, the logarithmic joint distribution of the first significant digits is defined by:

$$P(D_b^1 = d_1, \ldots, D_b^k = d_k) = log_b[1 + (\sum_{i=1}^{k} b^{k-i} d_i)^{-1}]$$

$\forall k \in N$; $d_1 \in \{1,\ldots, b-1\}$ and $d_i \in \{0,\ldots, b-1\}$, i=2,…,k.

(Its arguments continue to other digits and joint distributions of the digits.)

Example 3.2.2. taking the number 999889 in the formula above, we obtain:

$$P(D_1 = 9, D_2 = 9, D_3 = 9, D_4 = 8, D_5 = 8, D_6 = 9) = 4.34342037 \times 10^{-7}$$

The probability will be very low with a result close to 0. In the light of this, the worst fit of any data would be if the numbers were made up only of nines. This extreme case would occur if all significands are equal to $10^{(1-\varepsilon)}$,where ε denotes a very tiny quantity >0 [26].

3.3. Invariances

3.3.1. Scale invariance

Benford saw that the set of natural integers with one as first digit has no asymptotic natural frequency, as:

Definition 3.3.1. The limit below does not exist.

$$\lim \frac{1}{n} \quad with \quad \{i \in \mathbb{N} | i \leq n \quad and \quad d_1(i) = 1\}$$

Roger S. Pinkham [27], an American mathematician, offered an answer to this problem in 1961. He claimed that " if there was some Law governing digital distributions, this Law should be invariant by scaling". He discovered this property of the NBL which characterizes it, but the problem here is related to the universality of this reasoning.

Example 3.3.2. if we multiply our data by an arbitrary constant c> 0 (it must be positive as all the entries in a dataset that conforms NBL), e.g., c= ¾, and as the probability of the significand in [a, b] ⊂ [1, 10) is log(b/a), assuming base 10, we compute the probability that numbers have as leading digit 1, where we obtain the interval [3/4, 30/40). If we divide this interval by the parts containing 1 as the first digit we have that we can find 1 as first digit in the interval [1,2) and applying the formula said before, the probability of the interval is: log(2/1), which is the NB probability of the first digit, so we check how the distribution of the first significant digits remains unchanged. This is just what we do when we apply conversions in different units of measurement [4].

Property 3.3.3. ∀ n ∈ ℕ, ∀ d_1 ∈ {1,…,9}, ∀ d_k ∈ {0,…,9} with k ≥ 2 and ∀ c>0,

$$P(x : D_k(cx) = d_k) = P(x : D_k(x) = d_k) \quad \forall k = 1, \ldots, n$$

The probability distribution of a random variable x is independent of the measurement scale. If the data show deviations with NBL, these deviations shall

be present in all measurements. Therefore, our data must be measurable in different scales like Dollars/ Euros, miles/ km…

3.3.2. Base invariance

The values of the majority of the data do not usually depend on the base used. Hill [15] established the relationship between base-invariance and the uniformity of the mantissas of the logarithms [1].

Following the same idea than in previous section:

Property 3.3.4. X is a positive continuous random variable and it has the same significand distribution for two different bases b_1 and b_2 ⇒ X is NB for both bases [32].

Scale- invariance implies base-invariance, but not conversely.

3.3.3. Sum-invariance

As an Allaart's correction of what Nigrini observed first [20] we have: " A distribution is sum-invariant if for any natural number k, the sum of the significands of all entries starting with a fixed k-tuple of leading significant digits is the same as that for any other k-tuple". Sum invariance property is a characterization of the logarithmic Law [2].

Infinite datasets can obey NBL exactly. However, Nigrini observed first that real data, we refer to non-infinite datasets, approximately follows NBL and the sum of the significand of its entries with first digit 1 is similar as the sum with first digit 2 and so on [4].

Property 3.3.5. The sums of significands entries with $(D_1, \ldots, D_k)= (d_1, \ldots, d_k)$ are approximately equal ∀ tuples $(d_1, \ldots, d_k)$ of a fixed length k.

The Benford distribution is the only sum-invariant distribution, i.e. datasets of mixtures of other datasets.

3.3.4. Multiplication and division

Property 3.3.6. Y is a random variable that conforms to NBL and X is an arbitrary random variable with a continuous density ⇒ XY conforms to NBL.

Property 3.3.7. Y is a random variable ≠0 that conforms to NBL and X is an arbitrary random variable ≠0 with a continuous density ⇒ X/Y and Y/X conform to NBL.

The NB's distribution is invariant under inversion and multiplication by any distribution [20]. A sequence of products of random variables is very likely to converge to NBL, i.e. datasets from long series of computations.

3.3.5. Addition and subtraction

Related to addition the principle of invariance is not so trivial. R.W. Hamming [13], an American mathematician, has proved this property as other authors. He proved that the distribution of a sum of random variables converges to NBL. Empirically, it works quite well.

3.4. Sequences of interest

3.4.1. NBL sequences

Definition 3.4.1. A sequence $x_n = (x_1, x_2, ...)$ of real numbers is a NB sequence in base 10 if as $N \rightarrow \infty$, the limiting proportion of indices $n \leq N$ for which x_n has first significant digit d is [4]:

$$\lim_{N\to\infty}\left(\frac{\#\{1 \leq n \leq N : D_i(x_n) = d_i \quad \forall i = 1, \ldots, k\}}{N}\right) = log[1 + (\sum_{i=1}^{k} 10^{k-i} d_i)^{-1}]$$

$\forall\, n \in \mathbb{N}$, $\forall\, d_1 \in \{1,\ldots,9\}$, $\forall\, d_k \in \{0,\ldots,9\}$ with $k \geq 2$

3.4.2. Geometric sequence

The second part of Benford's paper was titled "The Geometric Basis of the Law", which explains a process with a constant growth rate where more time is spent at lower digits than at higher ones [4]. He associated the pattern of the digits with a geometric progression (as Raimi [29] did in 1976) where one of its main properties is that it is memoryless, the probability of an event does not depend on previous ones, and thus some of the best fits were for data whose numbers are not related to each other [20].

The geometric foundation of NBL means that a dataset will have NB-like properties if the ordered records closely approximate a geometric sequence[26]. The classical sequence 2^n also conforms to NBL.

3.4.3. Other Sequences.

NB sequences also include: n!, n^n, *Fibonacci* sequence and Lucas numbers (which follow a similar pattern to the Fibonacci sequence). Its conformity to the law improves as N increases, so do all sequences with similar recurring patterns. The numbers generated by the 3n+1 series, the Collatz conjecture, where n is any positive integer, also show NB tendencies. Nevertheless, these are dominated by the ending numbers in the sequence.

As curiosity, Zipf's Law is based on the number of times words are used in papers. With NBL, we would need a greater number of lower digits, but under Zipf's Law the trend is towards the most-often used words. Empirically, as it is obvious, data that conform to NBL do not conform to Zipf's Law, and vice-versa. [26]

To conclude, there are more sequences in the world that do not conform to NBL than do conform to NB, e.g. The sequence of primes does not conform to NBL.

3.5. Relation with the main distributions

Now, we are going to show the probability of occurrence of each digit, except zero, in different distributions and taking also into consideration the simulations' results by Formann [12] which are of special interest in order to sum them up.

We generate randomly in RStudio some selected distributions with n=100,000 simulations which correspond to: Cauchy(0, 1), Exponential(1), Normal(0,1), Chi-square with 1 degree of freedom, Uniform(0,1), Normal(5,1) and Chi-square with 100 degrees of freedom, placed in a particular order.

Thanks to the table below of the chosen distributions in the first four rows, we see that the digit one is the most frequent one, always exceeding the 30%, it appears as the leading significant digit. By contrast, we easily appreciate the progressive decrease in the percentages where the number nine appears less than 6%, so the results are close to NBL.

On the contrary, in the 5th row appears the results of a random generation of the U(0,1), where we obtain a very different output. All the digits occur with an equal frequency, it will take the value 0.0111 (1/90) approximately and depending on the n, so as we can see each occur about 11% of the time and this is because the uniform distribution and NBL are incompatible. However, what we know is that some mathematical operations, when performed on the uniform distribution, end up giving us NBL at the limit and Formann's paper also proves that its ratio fits better the law.

As we are working with the line of positive numbers, distributions like Cauchy(0,1) and N(0,1) seem to fit well to the law, but if we change its location/mean parameters things change (6th row). This is why symmetric distributions and those whose density increases as the value of its random variable are bad examples for NBL.

Not that trivial we find that distributions of the ratio of two random variables fit better than the distributions of a single random variable, e.g. the F-distribution, as ratio of two chi-squares, fits better than does chi-square. As we increase the df of a chi-square the fit will get worse because it approaches a normal distribution. (7th row)

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
C(0,1)	30.967	16.842	12.040	9.437	7.903	6.690	5.908	5.372	4.840
Exp(1)	32.993	17.485	11.405	8.447	7.232	6.351	5.798	5.368	4.921
N(0,1)	35.932	12.910	8.619	8.124	7.647	7.531	6.970	6.390	5.876
χ^2_1	30.586	18.347	12.509	9.314	7.701	6.394	5.759	4.984	4.406
U(0,1)	11.273	11.001	11.102	11.052	11.236	11.208	11.092	11.098	10.938
N(5,1)	0.129	2.068	13.560	34.409	34.036	13.493	2.154	0.149	0.002
χ^2_{100}	47.809	0.00	0.00	0.001	0.053	1.019	6.101	17.595	27.422

Table 3.1. *Probability of occurrence of each digit following them different distributions obtained with Rstudio.*

Therefore, the validity of the NBL requires that the frequency of small 'natural' numbers have to be predominant. The most common probability distributions are not exactly NB. However, some parametrized families are close for some specific values.

3.6. Convergence

If sets of numbers $x_1, \ldots, x_N$ are independent continuous random variables with densities $f_1, \ldots, f_N$ as $N \to \infty$, for many choices of densities, the distributions of the products converges to NBL and seem to be closer to NBL than its operands.

Definition 3.6.1. As $n \to \infty$ the distribution of the leading digits of X_n converges to NBL, and further that if X_1 is NB then X_n is also NB.

NBL is more robust than we might imagine, not all numbers will conform to NB's distribution, but if distributions are randomly selected and we take random samples from them, then the frequency of digits of this combined set of distributions will converge to NBL even if the separate distributions deviate from it [14].

3.6.1. Convergence to the uniform distribution

We might also remark that NBL is only useful for the first digits because the distribution of the k^{th} digit tends to U(0,9) exponentially fast as k increases [20] [9].

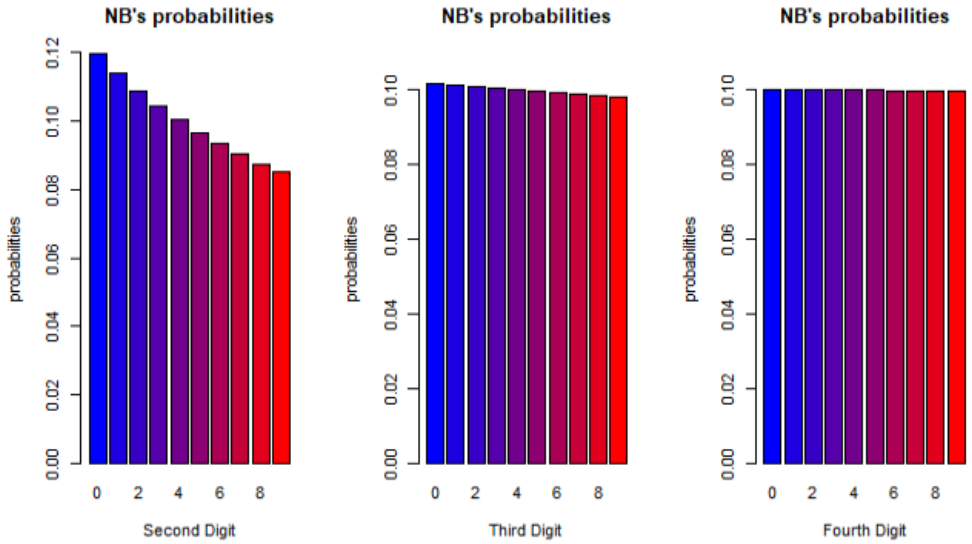

Figure 3.2. *Plot for second, third and fourth significant digits respectively showing their probabilities.*

From the fifth significant digit onwards the difference can barely be seen, if we take a few decimals it is equal to the graph of the density function of a U(0,9).

3.7. Bayesian approach

Classical methods, such as the chi-square test (See Section 4), could show problems in a NB's analysis. Because of this methods' weakness we can attend to Bayesian ones because datasets that reject these classical tests could still have a good fit to NBL.

In Bayesian statistics, it is essential the use of an initial distribution or prior, among others, although it is a hard task to explain its selection. The principle of insufficient reason establishes that if there is no information to differentiate between values other than its parameter θ, the same probability must be given to them. The principle implies a uniform prior for θ, where if the θ support is infinite, the prior distribution will be improper. In 1812, Laplace proposed it and it received many criticisms as the uniform distribution is not invariant in case of transformations. This means that it is not adequate for NBL, as we know one of its main properties is that it is scale-invariant.

In this way, Sir Harold Jeffreys whose considerations on the transformations of a problem had their beginnings, in order to be able to obtain non-informative priors, introduced what is known as the Jeffrey's prior and it is not affected by restrictions

on the parameter space, it is reparametrization-invariant. Moreover, it is the unique prior (on the positive real numbers) that is scale-invariant, so it is suitable for NBL [5], pp.85–86.

4. Empirical framework

4.1. Conditions for conformity

As we already know, many datasets satisfy NBL and within them, we highlight those where numbers represent sizes of facts or events, where they grow exponentially or emerge from multiplicative fluctuations or even, some well-known infinite integer sequences (like Lucas numbers) and the mixture of these datasets.

NBL also needs data that are not entirely random or highly conditioned. For example, the result of the lottery is completely random, each number has the same probability of appearing so it is not suitable for NBL.

Other data that do not conform to the Law are the numbers coming from evaluating functions such as: x^2, $x^{1/2}$ or 1/x.

Preferably, more than 1000 numbers should conform to this kind of datasets in order to avoid greater deviations from the NB's proportions and no minimum or maximum values should be incorporated into the data. Some suitable examples could be : death rates, diameter of planets, accounting transactions, stock prices, etc.

If our dataset does not conform this Law, this may be due to cases with narrow intervals for the data observations, numbers used as ids or labels, e.g. phone numbers or additive fluctuations instead of multiplicative ones in time series analysis, and the data should also have a larger amount of small numbers than higher ones, which is intuitive so that the data are not clustered around its mean value [26].

4.2. Statistical methods

We select some of the options provided by Rstudio libraries such as "benford.analysis" [7], which provide tools to validate data with using NBL, "BeyondBenford" [6], which compares the goodness-of-fit of NBL in a dataset and "BenfordTests" [10], which gives statistical tests for determining if a dataset conforms to NBL. These options include an analysis for the first and first-two significant digits.

4.2.1. Mantissa ARC Test.

We know mantissae would be uniform over [0,1) and as a proof to test it, the Mantissa ARC Test (MAT) was proposed. If the mantissae of a set of numbers are uniformly distributed on the circumference of a unit circle, the centre of the mass is at (0, 0) and the represented points on its circumference are equally spaced on it, in other cases it will be at the distance of L2, where its length is a measure of non-uniformity and if it is too large the data will not conform to NBL [1]. The centre of mass is called the mean vector, where its x and y coordinates are :

$$x = cos(2\pi\, m(a)) \qquad y = sin(2\pi\, m(a))$$

with angle 2πm(a) where m(a) is referred to its mantissa and a the number . Then, we average each component to obtain the mean vector and compute the distance L2, which is the squared euclidean distance,where $L2 = (mean(x))^2 + (mean(y))^2$

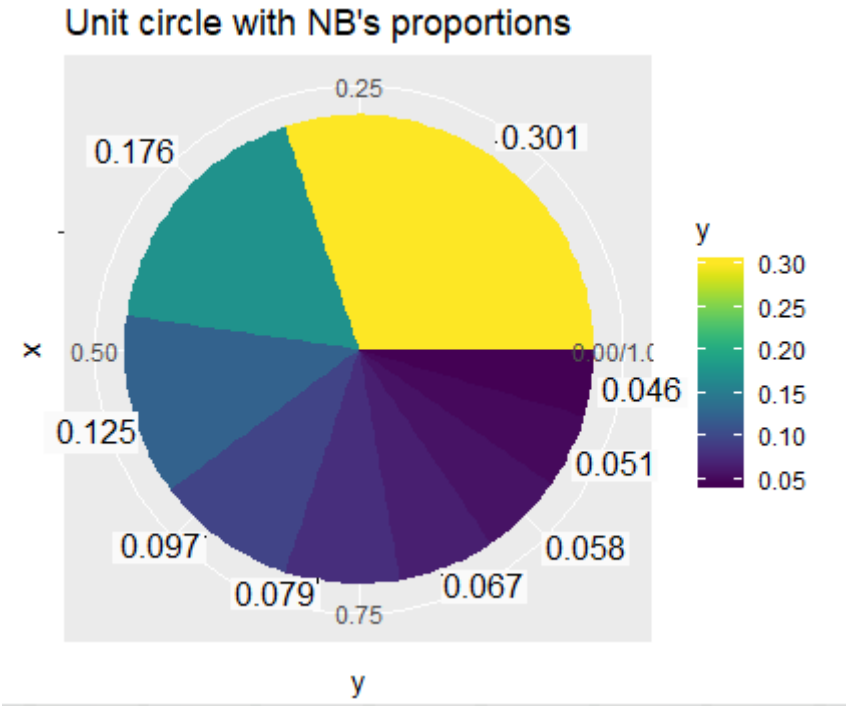

Figure 4.1. *Pie chart of the unit circle with the different probabilities of NBL.*

Each portion represents the probability of the first significant digits from 1 to 9 counter-clockwise. The points must lie on its circumference and depending on the section of the arc they are found, they would have one leading digit or another, e.g., a number whose first digit is 1, its mantissa will be in any point of the arc of the first portion, from 0 to 0.301. As mantissae are centred on (0,0), a mantissa= 0 will be on the position (0,1), a mantissa= 0.25 on (1,0), a mantissa= 0.5 on (-1,0) and a mantissa= 0.75 on (-1,-1).

The proposed hypothesis shall be:

H0: The mantissa of the observed distribution is uniformly distributed.

Under this hypothesis, the probability distribution of the size of the mean vector is precisely calculable. The larger L2, the more likely the null hypothesis H0 is to be rejected.

The distribution of the squared length of a two-dimensional normally distributed vector of length 1 is the chi-square distribution with two degrees of freedom, which is an exponential distribution, so in Rstudio's code [7] the p-value is calculated as:

p-value= exp(-(L2)*N), where N is the length of the dataset.

However it is a very sensitive test and its validity depends on whether the data are sufficiently dense in the unit circle.

4.2.2. Significand analysis

This test calculates confidence intervals using the Normal approximation for each digit only if N is large, taking as parameters:

$$\mu_i = np_i \quad and \quad \sigma 2 = np_i(1 - p_i)$$

where i identifies the digit, for the first digit $d_{i=1} \in \{1,\ldots,9\}$ and for the first-two digits $d_{i=1}d_{i=2} \in \{10,\ldots,99\}$, n is the length of the dataset and p_i the proportion of each digit.

Its graphical output in RStudio also gives the frequencies of the first or first-two digits and the p-value for investigating each individual significant digit [10].

4.2.3. Mean Absolute Deviation

The mean absolute deviation (MAD) test is a measure of conformity to NBL where the number of records, N, is not used and there are no objective critical values [7]. It is calculated as:

$$MAD = \frac{\sum_{i=1}^{K} |Actual\,Proportion(i) - Expected\,Proportion(i)|}{K}$$

where K represents the number of bins.

The higher the MAD, the larger the average difference between the actual and expected proportions.

Following Nigrini's suggestions [26], a MAD ≤ 0.0012 implies close conformity, a MAD ∈ [0.0012, 0.0018] as acceptable conformity , a MAD ∈ [0.0018, 0.0022] as marginally acceptable and a MAD≥ 0.0022 presents non conformity.

4.2.4. Chi-Square Test

The chi-square (χ2) test performs a goodness of fit test for the first and first-two digits [7].

Analysing the first digit , we compare the frequencies of the first significant digits of the chosen variable in a dataset with NB's frequencies with the same length N, so it compares these frequencies one by one and prints the corresponding plot. Of course, for finding NB's frequencies we need to use the first digit distribution formula (Definition 3.1.1) and then multiply it by N.

For the first-two digits, it follows the same principle as the previous explanation but it uses the first-two significant digits and Definition 3.1.5. It is more useful for finding biases in data and at detecting invented numbers. A weak fit to NBL usually suggests risk of errors, fraud or that the data are not suitable to conform NBL.

Then, since we have calculated the parameters needed in the formula below:

$$\chi^2 = N \times \sum_{j=1}^{K} \frac{(\hat{\theta}_j - f(j))^2}{f(j)}$$

where K represents the number of bins, N, the number of observations, θj is the observed frequency of digit j, and f(j) is the frequency of digit j implied by Benford's law. The number of degrees of freedom will be K–1.

H0: the sample comes from the distribution with probability density function $\{f_i\}_{i=1...n}$

The problem is that it is drastically affected by sample size, as it increases the results could be unfavourable to NBL.

Aside from this, the real datasets used will never conform perfectly to this law, we should not focus only on the significance of p-values.

The chi-square and K-S tests evaluate all the digits at the same time. These tests only tolerate small deviations from NBL for large N. For this reason, the best solution could be the MAD test because it does not take into account the number of records [26].

4.3. Real datasets

4.3.1. Introduction

There have been plenty of publications about newly discovered NB's datasets and, as they are mainly carried on by physicists and computing scientists, we are going to put this into practice too. Among the elements of a list of datasets that follow this Law, we can find the list of constants and measurements in physics, long series of floating-point operations, financial and accounting data or population datasets (Census data). The latter cases will be the ones chosen in our research. In the second dataset proposed we will try to detect possible manipulations following the indications from a dataset already analysed, described in Nigrini's book [26] and the data are included in RStudio, data("sino.forest"), where we will show its main characteristics in order to learn how to detect fraud.

Consequently, our main aim here will be to guess if these real-world datasets conform or not to NBL, or even, if they present some fraudulent indications, mainly following Nigrini's instructions from his book and it is also worth noting RStudio software.

4.3.2. 1st Dataset: "*Municipal population*"

We have selected a dataset from the INE's website, the National Institute of Statistics (*https://www.ine.es)*, about socio-demographic data of the registered population on 1st January 2019. (The dataset is available at [17] → "Población empadronada a 1 de enero de 2019" → Población de **60 o más años** por sexo: **Mapa municipal y de secciones censales** → Descarga de ficheros).

We have depurated our data which are eventually made up of 8132 observations and only four variables, which include "Municipality", "Age", "Sex" and "Total", where in the latter we will put special interest as it indicates the number of inhabitants per municipality.

The data belong to a wide interval of values, corresponding to sizes of the aforesaid fact, where they do not have any relationship among them, in principle. The primary conditions seem to be met and our dataset will be referred, from now on, as *"Municipal population".*

In a brief summary of the dataset, we see how the median differs a lot from the mean but this is because our country lives in a painful paradox: Spanish population has increased by about 36% since 1975, but this increase in the population is not balanced, the population is more concentrated in big cities such as "Madrid" and "Barcelona" (as we have proved with a chi-square test for detecting outliers), whose number of inhabitants is too large in comparison with almost the rest of Spain, making this fact mean value much higher than the

median. For this reason, the density function of our data are right-tailed, putting more mass on small values, so it follows some of the conditions presented before to fit the NBL.

	Min.	1stQuant	Median	Mean	3rdQuant	Max.
Total	3	154	525	5784	2392	3266126

Table 4.1. *Summary of the variable Total from "Municipal population".*

In order to continue with the NBL's study, we need first to calculate the frequencies of the first significant digits as well as compare them with the expected values as follows:

First Digit	Frequency	First digit proportions	NB probabilities	Absolute Difference
1	2471	0.3039	0.301	0.0029
2	1464	0.1801	0.176	0.0041
3	1011	0.1243	0.125	0.0007
4	734	0.0903	0.097	0.0067
5	674	0.0829	0.079	0.0039
6	526	0.0647	0.067	0.0023
7	478	0.0588	0.058	0.0008
8	386	0.0475	0.051	0.0035
9	387	0.0476	0.046	0.0016

Table 4.2. *Frequencies, proportions and absolute differences from "Municipal population" compared with NB proportions, $d_1 \in \{1,\ldots,9\}$.*

The best adjusted observed frequencies are those for the digits d_1= 3 and d_1= 7, while d_1= 4 is the worst and therefore this digit will be the one that least fits the curve set according to NB's conditions. These differences are represented as:

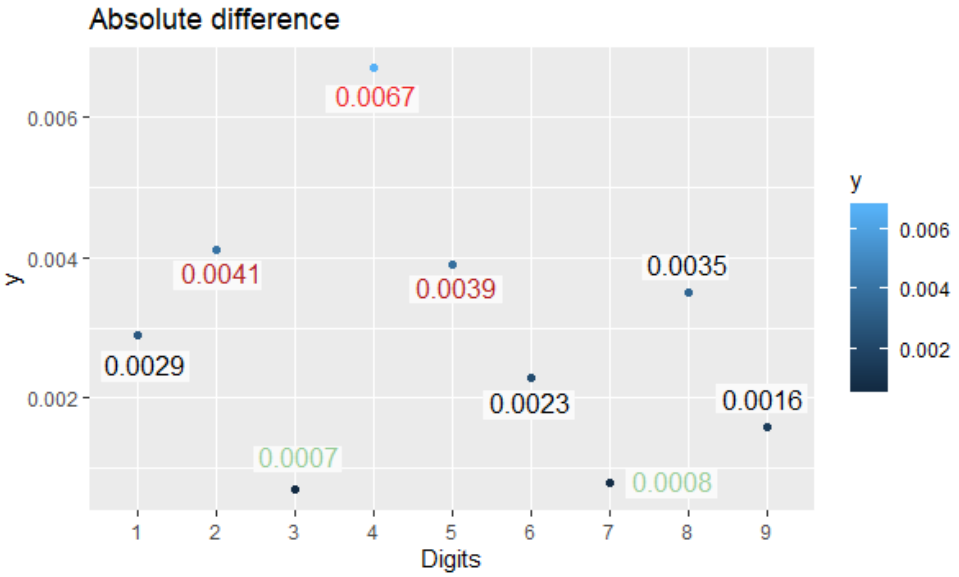

Figure 4.2. *Absolute differences of the first significant digits.*
- In red, the highest one corresponds to digit 4, in green the two-lowest ones.

It is easier to understand if we visualize it and thanks to the table below we can now plot a graph for the first digit, we see how our data follow NBL's behavioural patterns without anomalies and the differences of the observed and the expected data are not alarming at all.

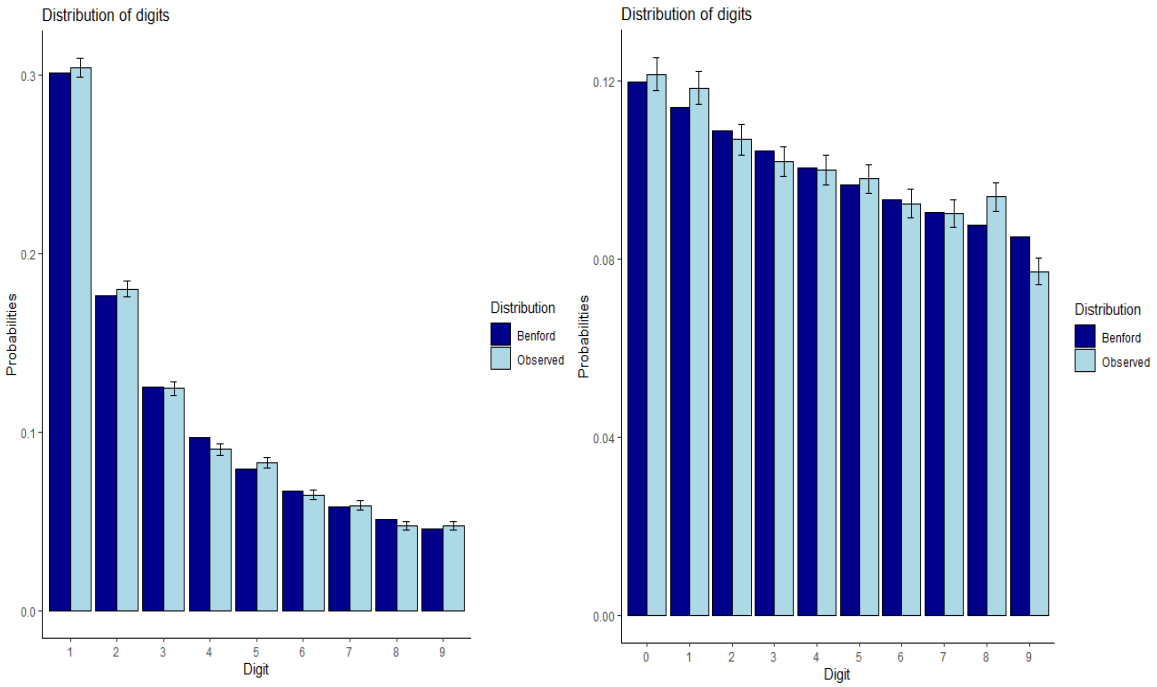

Figure 4.3. *The first plot shows the distribution of the first significant digits comparing the observed probabilities vs. NB's ones by using a red dotted line. The second one represents the same but using the second significant digit [6].*

Our study will focus on the first significant digits and the first-two, however we will graphically show the distribution of the second digit, in order to apply what we explained in Section 3.5. about the convergence to the uniform distribution. Comparing the graphs in Figure 4.3., the probabilities values in the second plot decrease in a more progressive way, trying to approach a uniform distribution as we increase the digit.

Digit	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Prob	0.121	0.118	0.107	0.102	0.099	0.098	0.092	0.09	0.094	0.077
NB	0.119	0.114	0.109	0.104	0.100	0.097	0.093	0.09	0.088	0.085

Table 4.3. *Probabilities' table of the second significant digit, $d_2 \in \{0,\ldots,9\}$.*

Nevertheless, the study does not finish here, just studying the first significant digit is not enough to assess NB's conformity. In the case of the first-two digits, the plot also shows that the data do have a tendency to follow NBL without any notable discrepancy.

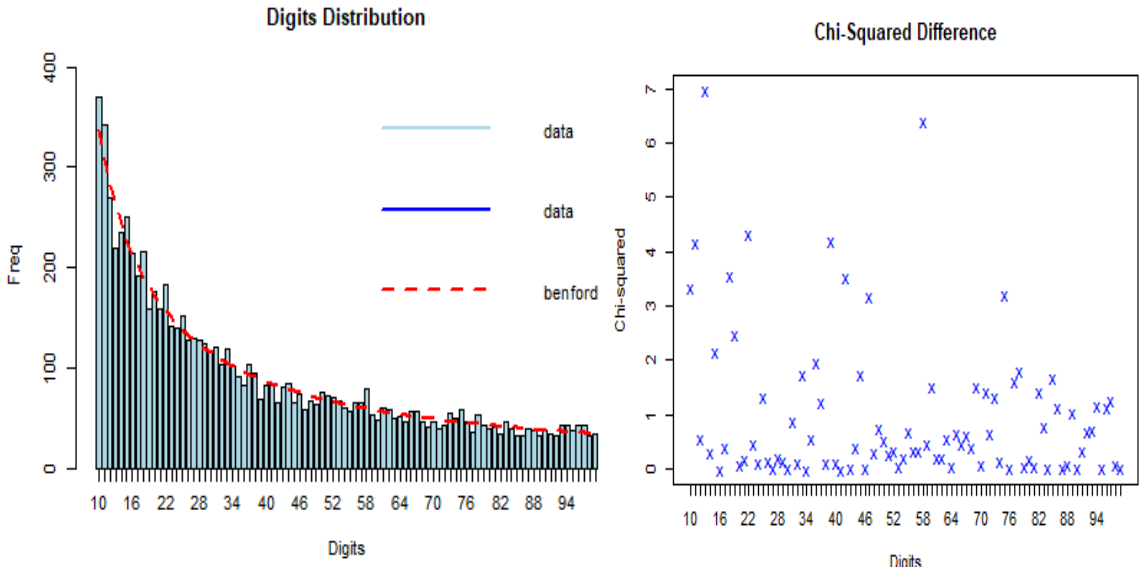

Figure 4.4. *Plot for the first-two digit distribution and chi-square differences respectively, where the second plot uses the squared differences of the frequencies [7].*

Note that for the chi-square test we use the frequencies instead of the proportions and for plotting its results the squared differences of the frequencies are needed.

Graphically, the Chi^2 differences show that digit 13 has the largest deviation. We complement this showing now the "suspects table", created as a data frame with the first-two digits of the absolute differences between expected and observed frequencies in decreasing order. It allows us to detect which digits present the greater deviation from the expected distribution.

Digits	Absolute difference
13	42.69
11	35.74
10	33.44
18	26.08
22	26.03

Table 4.4. *The five largest deviations from data "Municipal population".*

Digits 13 shows up as the first to appear in our table with an absolute difference of 42.69 as the highest one. By contrast, the lowest one corresponds to digits 22 and takes the value of 26.03. Therefore, 11 and 13 represent these suspicious numbers. We have found that the frequency of the first-two digits as 11 is 343 and for 13, 219.

Moreover, expanding the content and relating our results with what we know about the characteristics of the Spanish population, we should remark that Spain is the country with the largest population over 60 years but population under this age

tend to live in the most populated areas as we can see in the plot below, so the rest of our country consists almost of this aged population.

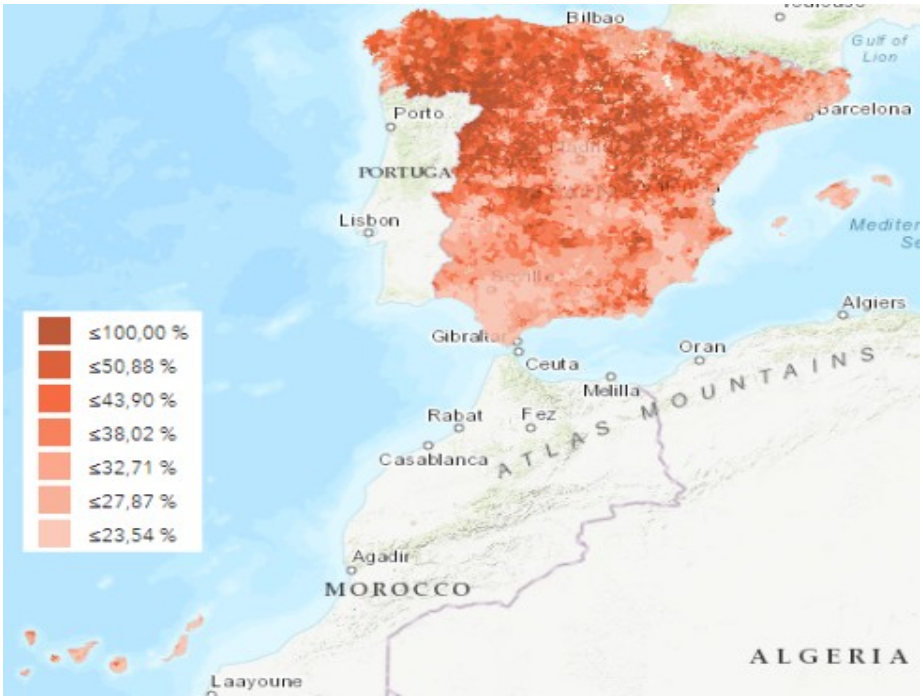

Figure 4.5. *Map of Spain with the municipality and census data levels. The legend shows the percentage of the population aged 60 or over [28].*

- On the one hand, our autonomous community leads this ranking, on the other hand, Madrid, Western Andalusia and the islands show the lowest percentage of population aged over 60 or more, meaning that a younger population predominates over these areas.

We have to look now at the **duplicated digits**, in our variable Total we have:

Total's values	***Duplicates***
56	*29*
33	*28*
57	*27*
106	*25*
22	*25*

Table 4.5. *The most duplicated first-two digits in descending order of appearance.*

The number 56 appears duplicated 29 times and the 33, 28 times, where we highlight municipalities from Castilla y León, as we might expect, and more specifically we remark the provinces of Á*vila* and *Burgos,* because the Total values obtained are low and our community is mostly unpopulated with aged population.

Some municipalities' examples which belong to Castilla La-Mancha, in the first case, and Castilla y León in the other cases are:

Municipality	***Total***
Villa de Ves	56
Aguas Cándidas	56
Villar de corneja	33
Hoyos del Collado	33

Table 4.6. *Examples of Municipalities with the first-two most duplicated values.*

Following the previous explanations, these low numbers of inhabitants appear more often than the rest indicating a low rate of usually aged population which is concentrated in certain areas as it can be seen in the map.

We can perform a Significand analysis [10]**,** where for the first digit, the significant p-values at 5% correspond only to the digit 4, p-value$_4$= 0.0430, so close to the significance level. If we remember, it was the digit with the largest deviation. For the first-two digits, looking at the p-values for each digit $d_1d_2 \in$ {10,...,99}, the significant ones correspond to:

d_1d_2	**frequency**	***p-value***
11	343	0.03764932
13	219	0.00730326
39	70	0.03906579
58	80	0.01119117

Table 4.7. *Frequencies and significant p-values of a significand analysis for the first-two digits.*

See the relationship with what we have commented before, digits 11 and 13 were already highlighted as the first-two digits with largest deviations. The p-values related to d_1d_2=47 and 75 are close to the significance level.
As regards the mathematical background of NBL, we are going to put into practice what we explained before.

Pinkham's proof [27] about the scale invariance property consists of the multiplication by a constant of a certain group of data where their distributions functions of mantissae will return the same distributions. We proceed the same way, by multiplying our variable Total by several different constants > 0 (we also need to calculate the mantissae where we add this constants instead of multiply them due to logarithm properties) and we perform a different Mantissa Arc Tests for each result of multiplying by the selected constant obtaining the same values:

The statistical result is: p= 6.7773e-05 and there is not strong enough evidence to reject the p-value= 0.5763 at a 5% significance level.

We conclude that our mantissae are uniformly distributed and they have taken these values:

Statistic	Expected value	Observed value
Mean	0.5	0.495
Standard deviation	0.288675	0.289826
Kurtosis	-1.2	-1.195
Skewness	0	0.014

Table 4.8. *Summary statistics of the log mantissa from data "Municipal population".*

where the expected values show the properties if the mantissae were perfectly distributed U(0,1) and the results show how they are really close to them.

Consequently, we have tested the scale invariance property and that our mantissae are uniformly distributed.

Finally, analysing the proposed statistics, for the first and first-two digits, we have:

	First Digit	First-two digit
MAD	0.00292753	0.0009472796
Conformity	Close	Close
χ^2 and df	**X^2**=9.5415, **df** =8	**X^2**=89.807, **df** =89
χ^2 p-value	0.2987	0.4561
K-S	0.61579	0.85595
K-S p-value	0.4846	0.3585

Table 4.9. *Main statistics for first and first-two digits.*

The **"MAD Conformity"** by Nigrini [26] says that our data have "Close Conformity" to the Law in both tests as its value is ≤ 0.0012.

As to the **Chi-Square test**, χ^2, where if we remember the degrees of freedom where K-1 (this is why in the first test applied they are only eight and in the second one, 89). H0 will be rejected if χ^2 is large. But as N is large, χ^2 also tends to be large. However, the p-values are not statistically significant, as they are higher than 0.05, so the null hypothesis is fulfilled.

The acceptance of H0 suggests that the data represent accurately the true behaviour of the variables measured and that the data follow the NB distribution. However, it is not warranted and its validity should be tested by contrasting results with other data quality metrics. It can be seen as a limit of the Neyman-Pearson theory.

The problem of goodness-of-fit in here is that it is a large dataset and the $\chi 2$ is not that precise. The simple distance to NBL that we have also used is more useful in these cases but it also presents an absence of accuracy and it depends on the data.

The **Kolmogorov-Smirnoff test** is also chosen here, it is a nonparametric test to determine the goodness of fit of two distributions, where one of them will be the

NB distribution. The p-values obtained for the first and first-two digits are not significant, so our data follow the NB distribution.

Although we discuss the results of the p-values during the study, it is worth noting real data will never conform perfectly to NBL so we might not focus on them.

To conclude, our dataset's choice has been guided by the fact that it is one of the types of datasets in which NBL has been conformed to, thanks to it we can also confirm that our dataset does follow NBL, where its results show a good fit to the Law despite the limitations. Because of this, the municipal population in Spain seems to have smaller digits which occur more often than greater ones.

4.3.2. Fraud detection guide

We are going to study a famous example of fraud detection described by Nigrini [26]. He denoted the corresponding section in his book as “Can’t see the forest for the trees”. Let us summarize his study and prove the results.

The data consist of 772 observations about financial statement numbers from Sino-Forest Corporation, which was one of the leading commercial forest plantation operators in China known for a research report in June 2011, due to this company reduced its stock price by almost 78% in a few days. The financial statement numbers were analysed and it was discovered that many information required have been omitted made with the intent to deceive. In 2012, the company filed for bankruptcy protection and also announced that it would be sold.

Studying the first-digits, the plot shows how the digits are close to NB’s curve, in red, looking at the MAD= 0.0065988135, it indicates “Acceptable conformity” and the largest deviation corresponds to digit 8. The first-two digits plot fits worse the NB’s curve and the MAD= 0.003463887 indicates “No conformity” to the Law.

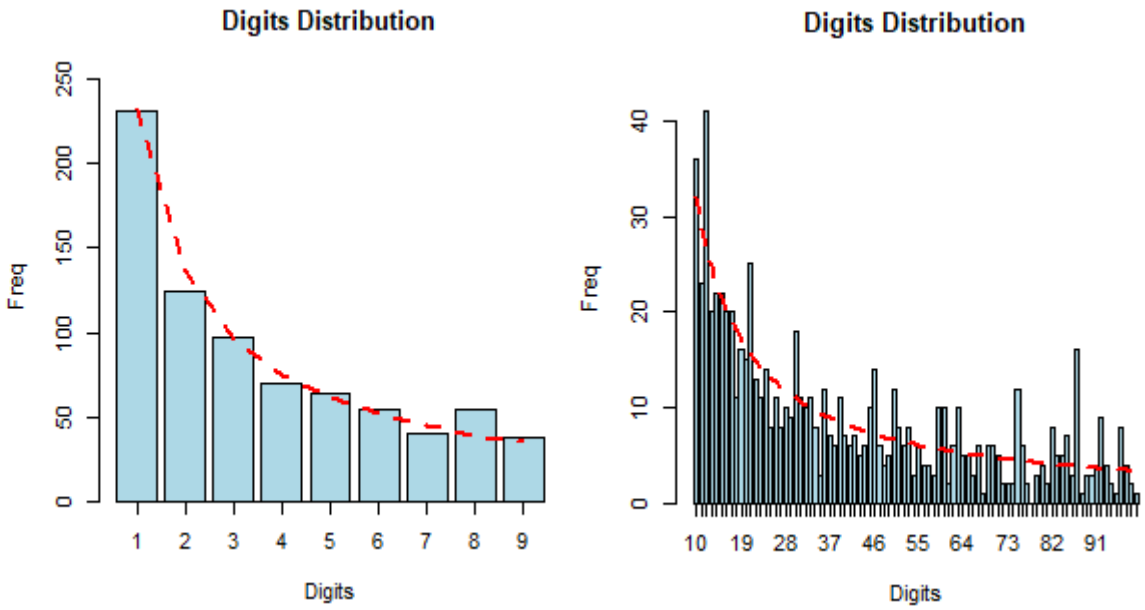

Figure 4.6. *Plot for the first-digit and first-two digits frequencies' respectively [7].*

Following with the first-two digits, the largest deviation is for 87, where Nigrini remarks that it is interesting that there is a spike at 87 but not at 88. If we look for duplicates we obtain several numbers which include those whose first-two digits are 87. The number 87670 is duplicated 5 times, 87670000, 3 times and all of them are related to the same debt instrument; 8756, which appears 3 times, is also involved in this transaction. However, duplicates do not imply fraud.

Although the dataset does not conform NBL, and this is an indicator of irregularities, deviations are not extreme. The spikes in the second plot showed repetitions of the same item, but only with the digit patterns we cannot conclude anything about the reported numbers, as there are simply few numbers in the dataset to accurately signal fraud and to achieve a good fit we need more than 1,000 entries [26]. However, some years later, in 2017, after the publication of Nigrini's book, it was released that Sino-Forest had committed fraud.

4.3.3. 2nd Dataset- "Stock Price(2012-2016)"

In our demographic dataset there is no reason to think about fraud, however in fields of economy things change, let us now study a dataset related to the five year daily stock price information, from 2012 to 2016, of the Japanese Company FastRetailing (Uniqlo), which is a public retail holding company.

The source of the dataset is Kaggle's website [19] and it is made up of data about this organization's purchases and sales of securities, where our main purpose is to identify suspicious data that will need further verification.

The chosen dataset consists of 1226 observations and 7 variables which include: *Date, open, High, Low, Close, Volume and Stock.Trading*. We are going to select the last variable, which collects information about the buying and selling of shares in the company and it has the widest range of values. As a brief description we have:

	Min.	1st Quan	Median	Mean	3rd Quan	Max.
Total	$3.966e^{9}$	$1.454e^{10}$	$2.154e^{10}$	$2.441e^{10}$	$3.016e^{10}$	$1.460e^{11}$

Table 4.10. *Summary of the variable Stock.Trading from "Training Data".*

First Digit	Frequency	First digit proportions	NB proportions	Absolute Difference
1	418	0.3409	0.301	0.0399
2	371	0.3026	0.176	0.1266
3	182	0.1485	0.125	0.0235
4	66	0.0538	0.097	0.0432
5	41	0.0334	0.079	0.0456
6	33	0.0269	0.067	0.0401
7	47	0.0383	0.058	0.0197
8	33	0.0269	0.051	0.0241
9	35	0.0286	0.046	0.0174

Table 4.11. *Frequencies, proportions and absolute differences from "Stock Price" compared with NB proportions, $d_1 \in \{1,\ldots,9\}$.*

We highlight the differences mainly between 1 and 2, being the last one more noticeable than usual. From d_1=4 onwards the differences begin to be smaller than NB's proportions and this fact will be reflected in Figure 4.10. where we will see how these proportions are not going to reach NB's curve.

Now we are going to focus on the first-two digits as they are more suitable for detecting fraud:

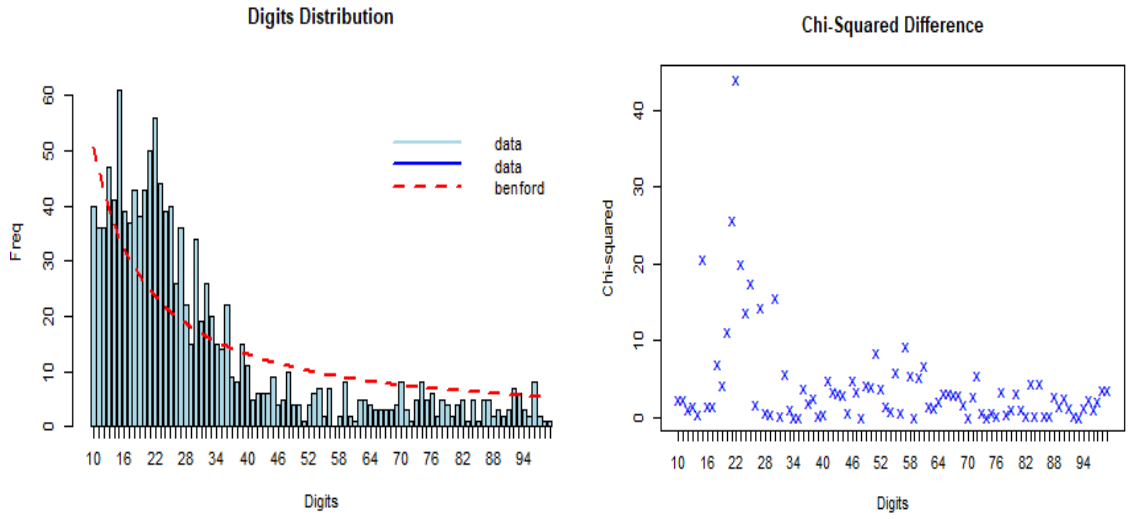

Figure 4.7. *Graph for the first-two digits and chi-squared test respectively [7].*

It is easy to appreciate how the first three dozens surpass the red dotted curve by far, but the majority of the values after d_1d_2=40 do not even reach NB's curve so we highlight overall the first part of the graph. Reinforcing what has been said with the help of the second plot, we can see that the five largest deviations in descending order come from the digits 22 , 15, 21, 23 and 25, where all of them belong to the interval [15, 23] , a narrow interval which may arouse suspicion about some kind of manipulations over this area.

Moreover, we find how digits one and two differ the most only looking at the first digit:

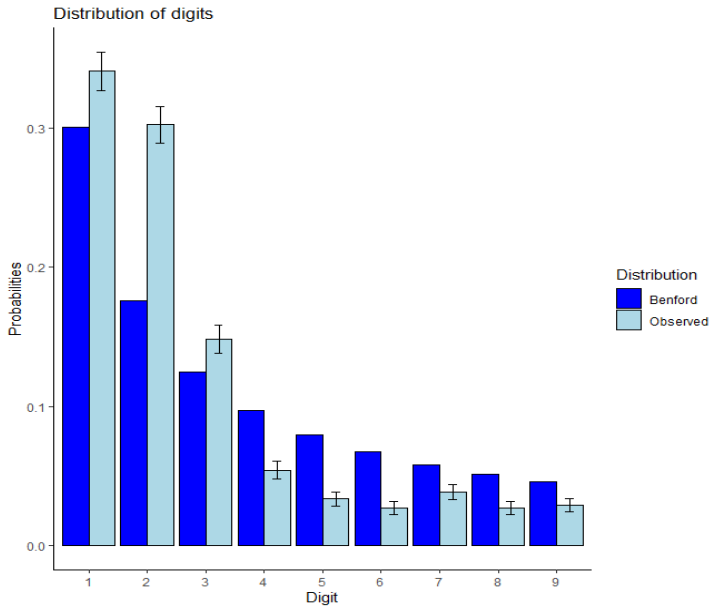

Figure 4.8. *Plot of the distribution of digits comparing the observed probabilities (light blue) vs. Benford's ones (blue), for the first and second digits respectively [6].*

Taking now the mantissae of this variable we obtain: (suppose the mantissae are uniformly distributed [0,1) and whose properties are the following)

Statistic	Expected value	Obtained value
Mean	0.5	0.420
Standard deviation	0.288675	0.246982
Kurtosis	-1.2	-0.396
Skewness	0	-0.562

Table 4.12. *Summary statistics of the mantissa.*

The obtained values differ much more than in the first dataset, the skewness now is negative. The mantissae does not seem to be perfectly distributed, although we will complete it with the Mantissa Arc Test, where its p-value is rejected so a significant difference does exist indicating that our suspicions have been confirmed.

Therefore, we can see how they tend to cluster in the plot.

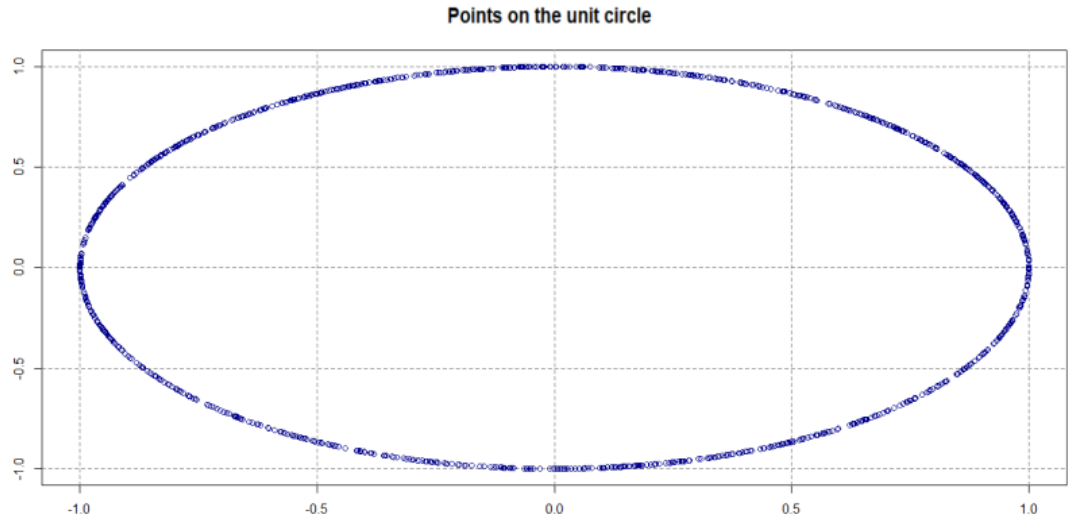

Figure 4.9. *Points on the unit circle to evaluate Mantissa Arc Test for the first-two digits. Note that the circumference is slightly flattened, like a geoid, in order to improve the vision of the points.*

The K-S and χ^2 tests are also rejected and the MAD=0.005100875 shows non conformity to the law, as it is ≥ 0.0022.

In closing, NBL is not conformed but the graphs seem to be similar to the curve that it produces, except for the first digits when analysing the first-two digits.
We have found several suspicious observations and we know that a weak fit to NBL means that there is a high risk the data contain anomalies.
Therefore, can we say that this is a case of fraud? Stock prices are referred to the current price that a share of stock is trading for, this can lead us to think that the company tended to change those prices in order to benefit themselves according to its own interests, but these are just mere assumptions. However, we should remember that the manipulation of stock prices can happen quite easily depending on the entity's trading power and this kind of event happens more often than we could imagine, but let us leave this to economic specialists as we do not know the company's economic history or trajectory.

4.4. Applications

The applications of the NBL are of great interest since from them it has been possible to develop much of the theory. These applications considerably increased with the arrival of computing, where huge masses of data need to be used [20]. The prevailing domains are:

4.4.1. Computation and Computer design

In the 70's, its use resurfaced with the arrival of computers where its application was in computation and in the design of computers. NBL took part mainly in the consideration of the distribution of computers' operands, in number representation. Significands are essential in floating-point arithmetic, where extremely large and small quantities of real numbers can be represented efficiently. This real numbers' codification requires a number of bits and depending on a certain pattern on the significand, a choice shall be made in order to optimize speed and storage.

The NBL can also be used to analyse different types of computational errors, e.g., round-off errors in the computation of products, to design algorithms in order to estimate their parameters' values as well as in the design of hardware and software, e.g., Hamming proposed a design problem which can be solved by applying NBL where he had to find a place to put the decimal point in his data in order to minimize the number of shifts in products.

4.4.2. Modelling

Another application used is mathematical modelling where NBL is useful to evaluate the predicted outcomes of mathematical models. In 1972, H. Varian [31] proposed the following idea: "If a certain set of values follow Benford's Law, then the models for the corresponding predictive values will also follow the Law".

Moreover, NBL has also provided a good diagnostic for selecting certain models in different types of researches such as an investigation about gas and condensed phases where NBL was the key for a good diagnostic in the selection of dynamical models or as a tool for controlling the quality of datasets produced by different measurement devices.

4.4.3. Fraud detection

Since in 1992 it was observed that financial data fit NBL [32], it has become of widespread use and this is why fraud detection is its main application. We need to define fraud detection as a fact typically committed by adding not suitable numbers, changing the real observations with a purpose, mainly lucrative, where

some of the dataset entries could be sums of others, duplicated, splitted into sums or even fabricated. In order to detect this kind of trap, NBL has become a popular tool.

Some interesting examples include:

- Electoral fraud: The presidential elections in Mexico in 2006, where certain districts were found to have anomalies in the votes' count because the distribution of the first digits, especially that of the second, did not conform to NBL. This led the authorities to a vote recount in order to verify the data provided by the Federal Electoral Institute [24].

- Tax return: Mark Nigrini also designed a computer program where an example of its applications was the analysis of Bill Clinton's tax return. Although it revealed that there were probably several rounding instead of exact figures, there were no indications of fraud. NBL's method, in turn, has some limitations in this field. Sometimes the figures cannot be given precisely and because of the rounding the first digit of a number, it can be modified [26].

- "White Collar" crimes: NBL is the clue to detect corporate crimes committed either by a corporation, company, individuals acting on behalf of a corporation, as in the case of Sino-Forest Corporation, or business entity, but studying this possible fraud deeply remains as a task concerning the relevant economic areas. Consequently, with NBL it is possible to detect a significant change in the reported figures by companies and/or individuals in consecutive years, where huge changes would indicate that something is going wrong.

- Currency impact: deviations from NBL were found after the introduction of the Euro in data from stock index returns, stock prices or consumer prices. Therefore, the application of NBL was a useful criterion for detecting anomalies in data [12].

- Clinical investigation: Recently, a research about COVID-19 has been carried out using NBL. It consisted of an analysis of the registered cases of people infected with the COVID-19 in 23 different countries. After the appliance of some tests, which include MAT, Chi-square test, etc. The results have shown that countries such as Italy, Portugal, Netherlands, UK, Denmark, Belgium and Chile are suspicious of data manipulation as they do not conform NB's patterns. However, it needs further study in order to verify it.
  There is a thing that draws our attention and it is that the sample was randomly selected, where 11 out of 23 countries were European ones, so almost 50%. However, after applying NBL, the possible cases of fraud belong in its majority to these European countries, except one, Chile. Nevertheless, authors consider that they must wait until the end of the pandemic to ensure what has been found and the source of the publication is not so reliable [18].

4.4.3.1. Machine learning

As NBL is a well-known procedure used in these fields, machine learning can be analysed for studying its behavioural patterns.

The task which concerns it, would be the selection of the classification rule which would be based on whether a dataset is fraudulent or not with a dichotomous variable that splits those chosen examples to fraud or legal ones. For example, the use of machine learning methods could be of interest if we had a dataset made up of information about other datasets at the same time. Thanks to this machine learning, we will improve forecasting to identify better the possible cases of fraud committed.

NBL also stands out in other fields such as diagnosis, detecting natural phenomena, as a pedagogical tool, verification of demographic models, in deciding the allocation of computer disk space, as a mean to identify problems with survey data, self-reported ratings, macroeconomic data quality, etc.

5. Conclusions

The purpose of this project is to summarize the most important aspects of NBL due to it is regaining its importance gradually as more findings are being made about it.
To close what we mentioned in the introduction, we have tried to convey the core concepts of the Law showing both mathematical and empirical frameworks:
On the one hand, in the mathematical background, we have seen how the first significant digits in nature does not follow a uniform distribution, how NBL meets different types of invariances and we have also studied its convergence and even some famous sequences and common distributions and how they follow NB's patterns.
On the other hand, in the empirical background, we have put NBL into practice with two real datasets concluding that the first dataset does conform to NBL but the second one not, perhaps due to possible fraudulent actions.
Accordingly, we have tried to regroup and check the most relevant information that involves this Law. However, there is something still incomplete with NBL, and maybe it is referred to the linkage of both frameworks, which keeps this law under constant review, or to the fact that it is not applicable to all datasets. In spite of this, NBL usually presents a good fit to empirical data because small objects or digits occur more often than do medium ones and they, in turn, occur more often than larger ones. It seems that these frequencies of occurrence follow an inverse function according to the objects' or numbers' sizes [12]. For this reason, inside datasets or other domains, NBL is usually conformed but we do not have a clear proof of it, e.g. the closing prices of stocks, the numbers of inhabitants, etc. We can conclude thanks to all of our work, that NB's distribution is non-uniform, with

smaller digits being more likely than larger ones, as can be seen empirically and data from any distribution will tend to follow NBL, as long as the distribution spans several orders of magnitude on the original scale and as long as the distribution is reasonably smooth.

12We want also to give credit to all the authors mentioned, but especially to Mark Nigrini, even though he is not a scientist, his work on NBL is noteworthy because of its ease of understanding and usefulness today which have been helpful for us to comprehend the law properly. Thanks to him, this law has regained its interest and has became more popular, being its use now almost indispensable in certain areas, especially in the field of the economy, where this law is growing more and more because of its utility in terms of auditing and fraud detection (due to NBL's application gives us the opportunity to know if a random dataset is fraudulent or not as deviations from NBL can be caused by irregularities or anomalies). Therefore, more countries are adding to the use of the Law, like Nigrini's native country, the USA, where NBL is free to apply.

6. Appendix

```
#Load packages
library(ggplot2)
library(dplyr)
library(stringr)
library(magrittr)
library(benford.analysis)
library(BeyondBenford)
library(BenfordTests)
library(sm)
library(outliers)
#colours
library(viridis)

########3.Mathematical justification.#########

#3.1.The first digit distribution
#Figure 3.1
#The distribution of the first significant digit
benfd1 <- log10(1+1/(1:9)) # log(1+1/(1:9), base = 10)
data.frame(benfd1)
#generation of a uniform(0,100)
N <- 1000
x <- runif(N, 0, 1)
freq <- fd1(x)
freq
#Plot
df <- data.frame(x = 1:9, y = benfd1)
ggplot(df, aes(x = factor(x), y = y) ) + geom_bar(stat = "identity", fill=heat.colors(9))
+
                                          xlab("First                    Digit")+ylab(NULL)
+geom_line(data=freq,aes(x=Var1,y=Freq,group=1),colour="black",size=2)+
  geom_point(data=freq,aes(x=Var1,y=Freq,group=1),colour="black",size=4,pch=2
3,bg="black")

#########################################################################
#3.5.Relation with the main distributions
#For explaining Formann's paper for the first digit
#Simulation  for the first digit
simd1<- function( distro, ..., n = 100000 ){
          digit <- distro( n, ... )
          digit <- as.character( digit[ digit > 0] )
           digit <- strsplit( digit, "" )
           first<- function( x )
             x[ ! x %in% c( "0", "." )][1]
```

```
          digit <- unlist( lapply( digit, first ) )
          table( digit ) / length( digit )
        }
#Cauchy(0,1)
simd1( rcauchy )
#Changing its parameters
simd1( rcauchy,5,1)

#Exponential(1)
simd1( rexp, rate = 1 )

#Normal(0,1)
simd1( rnorm,0,1)
#Changing its parameters
simd1( rnorm,5,1)

#Uniform (0,1)
simd1(runif)
#similar results U(0,100)
simd1(runif,0,100)

#Ch-sqare df=1
simd1(rchisq,df=1)
#worse results, chi-square df=100
simd1(rchisq,df=100)

#Other distributions:
simd1(rlogis)
simd1(rlnorm)
#F distribution
simd1(rf,df1=1,df2=1)
#worse results
simd1(rf,df1=100,df2=100)

#########################################################################
#3.6.Convergence
#First two-digits, from 10 to 99
par(mfrow=c(1,3))
benford1d2=log10(1+(1/((10:99))))
names(benford1d2) = 10:99
aux = 10:99 - floor(10:99/10)*10
benford2 = (data.frame(v1 = benford1d2,v2 = aux) %>% group_by(v2) %>%
summarise(v1 = sum(v1)))$v1
names(benford2) = 0:9
barplot(benford2, xlab = "Second Digit",
          ylim = c(0, .12),ylab="probabilities",col= colorRampPalette(c('blue', 'red'))
(10),main="NB's probabilities")
```

```
#We check both sums
sum(benfd1)#1
sum(benford1d2)#1

#we pass the probabilities directly
#third significant digit
x3=c(0.10178,0.10138,0.10097,0.10057,0.10018,0.09979,0.0994,0.09902,0.0986
4,0.09827)
barplot(x3, names.arg = 0:9, xlab = "Third Digit",
          ylim = c(0, .11),ylab="probabilities",col= colorRampPalette(c('blue', 'red'))
(10),main="NB's probabilities")
#fourth significant digit
x4=c(0.1002,0.1001,0.1001,0.1001,0.1,0.1,0.0999,0.0999,0.0999,0.0998)
barplot(x4, names.arg = 0:9,ylab="probabilities",xlab = "Fourth Digit",
            ylim = c(0, .11),col= colorRampPalette(c('blue', 'red'))(10),main="NB's
probabilities")

#Same probabilities for the fifth significant digit
#x5=c(0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1)
#barplot(x5, names.arg = 0:9, xlab = "Fifth Digit",
        #   ylim = c(0, .11),ylab="probabilities",col=col= colorRampPalette(c('blue',
'red'))(10),main="NB's probabilities")

#########################################################################

#4.2.1. Mantisa Arc Test (MAT)
#Plots for the explanation of MAT
bfp=c(0.301, 0.176, 0.125, 0.097, 0.079, 0.067, 0.058, 0.051, 0.046)
#con ggplot
ggplot(df, aes(x = "", y = y,fill=y,labels = bfp )) +
  geom_bar(width=1,stat = "identity")  +
  coord_polar("y", start=3*pi/2,direction=-1)+
  scale_color_viridis(discrete = FALSE, option = "C")+
  scale_fill_viridis(discrete = FALSE) +ggtitle("Unit circle with NB's proportions")

#Plot for the mantissa arc test in the first dataset: municipalitys population
plot(0,0,asp=1,type="n",ann=F, xlim = c(-.5, 0.5), ylim = c(-1,1))
mantisa=as.numeric(datosBenford$LogTotal)-Ent
t=2*pi*mantisa
abline(v = seq(-1, 1, 0.5), lty = 2, col = "gray50")
abline(h = seq(-1, 1, 0.5), lty = 2, col = "gray50")
x = cos(t)
y =sin(t)
points(x,y,col="black")

#Plot for the mantissa arc test in the third dataset: stockPrice because MAT is
rejected
plot(0,0,asp=1,type="n",ann=F, xlim = c(0, 0), ylim = c(0,0))
```

```
stockPrice$LogVolume=log(as.numeric(stockPrice$Volume))
Ent<-floor(as.numeric(stockPrice$LogVolume))
mantisa=as.numeric(stockPrice$LogVolume)-Ent
t=2*pi*mantisa
abline(v = seq(-1, 1, 0.5), lty = 2, col = "gray50")
abline(h = seq(-1, 1, 0.5), lty = 2, col = "gray50")
x = cos(t)
y =sin(t)
points(x,y,col="dark blue")

#Plot for the mantissa arc test in the second dataset: sino.forest
plot(0,0,asp=1,type="n",ann=F, xlim = c(0, 0), ylim = c(0,0),main="Mantissae on the unit circle")
sino.forest$Log=log(as.numeric(sino.forest$value))
Ent<-floor(as.numeric(sino.forest$Log))
mantisa=as.numeric(sino.forest$Log)-Ent
t=2*pi*mantisa
abline(v = seq(-1, 1, 0.5), lty = 2, col = "gray50")
abline(h = seq(-1, 1, 0.5), lty = 2, col = "gray50")
x = cos(t)
y = sin(t)
points(x,y,col="gray")
plot(mantisa,main="Points on the unit circle")

#########################################################################
#4.3.2. Municipal population

#Registered population at 1 January 2019, population aged 60 or over
#Corrected dataset due to the use of "ñ" and tildes in the original
#we took the rows for both genders
#We change the names
datosBenford1=read.csv2("C:/Users/Admin/Desktop/
Benford'sLaw.History,mathematical justification and
applications/Datasets/pob60ymas_mun.csv",sep=";")
names(datosBenford1)=c("Municipio","Edad","Sexo","Total")
datosBenford1$Edad=as.factor(datosBenford1$Edad)
levels(datosBenford1$Edad)=c("Poblacion de +60","Poblacion total")
datosBenford2=datosBenford1[datosBenford1$Sexo=="Ambos sexos",]
#For total age, the dataset conforms NBL. From datosBenford2
datosBenford4=datosBenford2[datosBenford2$Edad=="Poblacion total",]
datosBenford=datosBenford4[-c(1),]

head(datosBenford)
attach(datosBenford)
barplot(table(Total),main="Barplot for Total")
dens=density(Total)
plot(dens)
#For first digit proportions
```

```
prop <- Total %>%
  as.character() %>%
  str_extract(pattern = "[^0\\.]") %>%
  substr(1,1)
prop <- factor(prop, levels = 1:9)
table(prop) %>% prop.table()

#Brief study of our data
#Removing madrid obs. as the largest
datosBenford$num=1:nrow(datosBenford)
#related to outliers obs. num. 4369 and 883
datosBenfordOut=datosBenford[-c(4369,883),]
summary(datosBenfordOut$Total)
#Outliers
chisq.out.test(datosBenfordOut$Total)
#The plots remain similar
#We avoid this, coming back to the begginning -> datosBenford
#Density
hist(datosBenfordOut$Total)
sm.density(datosBenfordOut$Total)
#main statistics of ben
getBfd(ben)

#datosBenford=as.data.frame(datosBenford)
#No NA's
#Generates Benford objects
#For first two digits
ben=benford(datosBenford$Total,2)
ben
plot(ben)
#For first digit
ben1=benford(datosBenford$Total,1)
ben1
plot(ben1)
#plotting the differences between empirical frequencies proportios and NB ones.
x=c(0.0029,0.0041,0.0007,0.0067,0.0039,0.0023,0.0008,0.0035,0.0016)
absDif=data.frame(x = 1:9, y = x)
ggplot(data = absDif, mapping = aes(x = factor(x), y = y,col=heat.colors(9)))+
          geom_point()+geom_line()+labs(x="Digits")+ggtitle("Absolute difference")



#Scale invariance

mantissa(ben)
#To calculate the mantissas through the logs of the observations
datosBenford$LogTotal=log(as.numeric(datosBenford$Total))
```

```
sortLog=sort(datosBenford$LogTotal, decreasing = FALSE)

#Plot for ordered logs
old.par <- par(no.readonly = TRUE)
par(bg = "gray98")
par(col.axis = "black")
par(col.lab = "black")
par(font.lab = 4)
par(font.main = 4)
plot(sortLog,cex.lab=0.8,pch=24,bg="7",col = "honeydew4",main="Ordered logs of
total population", ylab="Log of total population", xlab="Rank",col.axis="dark gray")


#First way for finding the mantissa
Ent<-floor(as.numeric(datosBenford$LogTotal))
datosBenford$mantissa=as.numeric(datosBenford$LogTotal)-Ent

#Second way applying Nigrini's formula
N=8132
datosBenford$mantissa2= ((rank(datosBenford$mantissa)-1)/N) + (1/(2*N))

#Multiplying by a constant
datosBenford$mantissaChange=as.numeric(datosBenford$mantissa)+3.162
datosBenford$mantissaChange2=as.numeric(datosBenford$mantissa2)+3.162
datosBenford$Total2=as.numeric(datosBenford$Total)*3.162

scaleInv1=data.frame(datosBenford$Municipio,datosBenford$Edad,datosBenford$
Sexo,datosBenford$Total2,datosBenford$mantissaChange)
benford(scaleInv1$datosBenford.Total2)

#by 1.5
datosBenford$mantissaChange=as.numeric(datosBenford$mantissa)+1.5
datosBenford$mantissaChange2=as.numeric(datosBenford$mantissa2)+1.5
datosBenford$Total2=as.numeric(datosBenford$Total)*1.5

scaleInv2=data.frame(datosBenford$Municipio,datosBenford$Edad,datosBenford$
Sexo,datosBenford$Total2,datosBenford$mantissaChange)
benford(scaleInv2$datosBenford.Total2)

#by 9
datosBenford$mantissaChange=as.numeric(datosBenford$mantissa)+9
datosBenford$mantissaChange2=as.numeric(datosBenford$mantissa2)+9
datosBenford$Total2=as.numeric(datosBenford$Total)*9

scaleInv3=data.frame(datosBenford$Municipio,datosBenford$Edad,datosBenford$
Sexo,datosBenford$Total2,datosBenford$mantissaChange)
benford(scaleInv3$datosBenford.Total2)
```

```
#####################################
#For getting duplicates
getDuplicates(ben, datosBenford)
duplicatesTable(ben)
getDuplicates(ben1, datosBenford)
duplicatesTable(ben1)
#####################################
#suspicious data
suspects <- getSuspects(ben, datosBenford)
suspects
suspectsTable(ben, by="absolute.diff")
#####################################
#extract the first digit
datosBenford$first_digit=substr(Total,1,1)
#extract the second digit
datosBenford$second_digit=substr(Total,1,2)
attach(datosBenford)
#Filter amounts starting with the suspected digit:
suspects_orders1 <- subset(datosBenford,first_digit== 1)
suspects_orders1
summary(suspects_orders1)
suspects_orders2 <- subset(datosBenford,second_digit== 13)
suspects_orders2
summary(suspects_orders2)
#Filter amounts starting with the suspected digit:
suspects_orders1
summary(suspects_orders1)
suspects_orders2 <- subset(datosBenford,second_digit== 11)
suspects_orders2
summary(suspects_orders2)
######################
#Summary of stats
ben$bfd
summary(ben$bfd)
######################
#dat.distr(datosBenford$Total), for first and first-two digits
digit.distr(datosBenford$Total,  dig=1,  mod="ben",  No.sd=1,  Sd.pr=1,col=c("dark
blue","light blue"))
digit.distr(datosBenford$Total,  dig=2,  mod="ben",  No.sd=1,  Sd.pr=1,col=c("dark
blue","light blue"))
## Measure of Benford's Law goodness of fit code in benford.analysis [7]
#Code behind Chi-square test in [7]
#squared.diff=((empirical.distribution$dist.freq-benford.dist.freq)^2)/
#benford.dist.freq
#chisq=sum(squared.diff)#the chi-squared diff. between data and NB freq.
#chisq.pvalue=pchisq(chisq,df,lower.tail=F)
obs.numb.dig(Total, dig=2)
```

```
obs.numb.dig(Total, dig=1)
######################
#For different tests, Beyond Tests
help("BenfordTests")
#Kolmorov-Smirnov test
ks.benftest(x = Total, digits = 2, pvalmethod = "simulate", pvalsims = 10000)
#Significand analysis
signifd.analysis(Total,graphical_analysis=TRUE,freq=TRUE,tick_col="red",
            ci_col="light blue",ci_lines=c(0.05))
#The same for the first two digits
signifd.analysis(Total,digits=2,graphical_analysis=TRUE,alphas=0.05,freq=TRUE,t
ick_col="red",              ci_col="pink")


################################Datasets about fraud####################
#4.3.3. Fraud detection guide
        data("sino.forest")
        benForest1=benford(sino.forest$value,1)
        benForest1
        plot(benForest1)
        benForest=benford(sino.forest$value,2)
        benForest
        plot(benForest)
        #Suspects
        suspectsTable(benForest)
        getSuspects(benForest, sino.forest)
        #Extract the first digits
        left=function(string,char){
          substr(string,1,char)
        }
        sino.forest$d11=left(sino.forest$value,1)
        #frequencies of each digit
        obs.numb.dig(sino.forest$value,2)
        obs.numb.dig(sino.forest$value,1)

        #For getting duplicates
        getDuplicates(benForest, sino.forest)
        duplicatesTable(benForest)

#########################################################################
#4.3.4. Training Data (2012-2016 Stock Price)
        stockPrice=read.csv("C:/Users/Admin/Desktop/
        Benford'sLaw.History,mathematical          justification          and
        applications/Datasets/stocks.csv",sep=",")
        attach(stockPrice)
        #For first digit
        prop <- Stock.Trading %>%
          as.character() %>%
```

```
  str_extract(pattern = "[^0\\.]") %>%
  substr(1,1)
prop <- factor(prop, levels = 1:9)
table(prop) %>% prop.table()

stockbf=benford(stockPrice$Stock.Trading,2)
stockbf
plot(stockbf)
#frequencies of each digit
obs.numb.dig(Stock.Trading, dig=1)
obs.numb.dig(Stock.Trading, dig=2)
#chi2 Test
chi2(Stock.Trading, dig=2, pval=1)
#Suspects
suspectsS<- getSuspects(stockbf, stockPrice)
suspectsS
suspectsTable(stockbf, by="absolute.diff")

#extract the first digit
stockPrice$first_digit=substr(Stock.Trading,1,1)
#extract the second digit
stockPrice$second_digit=substr(Stock.Trading,1,2)
attach(stockPrice)
#Filter amounts starting with the suspected digit:
suspects_orders1 <- subset(stockPrice,first_digit
                    == 1)
suspects_orders1
summary(suspects_orders1)
suspects_orders2 <- subset(stockPrice,second_digit
                    == 15)
suspects_orders2
summary(suspects_orders2)
#Filter amounts starting with the suspected digit:
suspects_orders1 <- subset(stockPrice,first_digit
                    == 2)
suspects_orders1
summary(suspects_orders1)
suspects_orders2 <- subset(stockPrice,second_digit
                    == 22)
suspects_orders2
summary(suspects_orders2)

#For getting duplicates
getDuplicates(stockbf, stockPrice)
duplicatesTable(stockbf)

#Digit distribution comparing observed prob. vs. Benford for first-two digits
```

```
digit.distr(Stock.Trading, dig=2, mod="ben", No.sd=1, Sd.pr=1,col=c("grey","pink"))
digit.distr(Stock.Trading, dig=1, mod="ben", No.sd=1, Sd.pr=1,col=c("blue","light blue"))

#Kolmogorov-Smirnov test
ks.benftest(x = Stock.Trading, digits = 2, pvalmethod = "simulate", pvalsims = 10000)
```

7. Bibliography

1. Alexander JC (2009). Remarks on the use of Benford's law. Working paper, Case Western Reserve University, Department of Mathematics and Cognitive Science.

2. Allaart PC (1997). An invariant Sum-Characterization on Benford's Law. Volume 34 Issue 1.

3. Benford F (1938). The Law of Anomalous numbers. Proc Am Phil Soc 78: 551–572.

4. Berger A, Hill TP (2011). A basic theory of Benford's Law. Probab Surv 8: 1–126. 10.1214/11-PS175

5. Berger JO (1985). "Prior information and subjective probability" in Statistical Decision Theory and Bayesian Analysis. Springer-Verlag: Springer Series in Statistics.

6. Blondeau Da Silva Stephane (2020). BeyondBenford: Compare the Goodness of Fit of Benford's and Blondeau Da Silva's Digit Distributions to a Given Dataset. R package version 1.4. https://CRAN.R-project.org/package=BeyondBenford

7. Carlos Cinelli (2018). benford.analysis: Benford Analysis for Data Validation and Forensic Analytics. R package version 0.1.5. https://CRAN.R-project.org/package=benford.analysis

8. Cirillo A (2016). RStudio for R Statistical Computing Cookbook. Birmingham, UK:Packt Publishing.

9. Diaconis, P., & Freeman, D. (1979). On rounding percentages. Journal of the American Statistical Society, 74(366), 359–364.

10. Dieter William Joenssen (2015). BenfordTests: Statistical Tests for Evaluating Conformity to Benford's Law. R package version 1.2.0. https://CRAN.R-project.org/package=BenfordTests

11. Fewster R M (2009) A Simple Explanation of Benford's Law. Amer. Statist. 63, 26–32.

12. Formann AK (2010). The Newcomb-Benford Law in Its relation to some common distributions. PLoS ONE 5: e10541 10.1371/journal.pone.0010541.

13. Hamming R (1970). On the distribution of numbers. Bell System Technical Journal, 49: 8. October *1970* pp 1609-1625.

14. Hill TP (1995). A Statistical derivation of significant digit-law. Stat Sci 10: 354–363.

15. Hill TP (1995). Base-Invariance implies Benford's Law. Proc Am Math Soc 123:887–895.

16. Hill TP (1995). The significant-digit phenomenon. Am Math Mon 102: 322–327.

17. Información estadística para el análisis del impacto de la crisis COVID-19. Datos sociodemográficos. (2020) INE *(URL:https://www.ine.es/covid/covid_sociodemo.htm)*

18. Isea R (2020). How Valid are the Reported Cases of People Infected with Covid-19 in the World?INTERNATIONAL JOURNAL OF CORONAVIRUSES. Vol-1 Issue 2 Pg. no.– 53

19. Ishii D (2017). Uniqlo (FastRetailing) Stock Price Prediction. Tokyo Stock Exchange Data (LightWeight CSV) in 2016 for Beginners. (URL:

https://www.kaggle.com/daiearth22/uniqlo-fastretailing-stock-price-prediction)

20. Jamain A (April-September 2001). Benford's Law Master's thesis, Department of Mathematics, Imperial College of London and ENSIMAG, London, UK, 2001. (URL:

http://www.math.ualberta.ca/~aberger/benford_bibliography/jamain_thesis01.pdf.)

21. Jang D et al. (2009). Chains of distributions, Hierarchical Bayesian models and Benford's Law. Journal of Algebra, Number Theory: Advances and Applications, volume 1, number 1 , 37–60.

22. Journal of Applied Probability Volume 41 (2004): Index - Volume 41 Issue 4.pp. 1250-1254

23. Miller S J (2015). Benford's Law: Theory and Applications. 1251–1252

24. Mebane WR (2008). The second-digit Benford's law test and recent American presidential elections. In: Alvarez, R. M., Hall, T. E., Hyde, S. D. (Eds.), Election Fraud: Detecting and Deterring Electoral Manipulation. Brookings Press, Washington, D. C., pp. 161–181.

25. Newcomb S (1881). Note on the frequency of use of the different digits in natural numbers. Am J Math 4: 39–40.

26. Nigrini MJ (2012). Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection. Hoboken, NJ.

27. Pinkham RS (1961). On the distribution of first significant digits. Ann Math Stat 32: 1223–1230.

28. Población empadronada a 1 de enero de 2019. Densidad de población: Mapa municipal (01-01-2020) INE.

(URL:

https://inespain.maps.arcgis.com/apps/webappviewer/index.htmlid=179c7bc5ac5e41b4a0237cf071f3b1b1)

29. Raimi R (1976). The first digit problem. Am Math Mon 83: 521–538.

30. Schatte P (1996). On Benford's law to variable base. Stat Probabil Lett 37: 391–397.

31. Varian HR (1972). Benford's Law. Amer. Statist. 25, 65–66.

32. Wojcik, M. R. 2013, Notes on scale-invariance and base-invariance for Benford's law, arXiv:1307.3620.

7. List of figures and tables