"Science Stays True Here"
Journal of Mathematics and Statistical Science (ISSN 2411-2518, USA), Vol.8 (2022), Issue 10, 51-58 | Science Signpost Publishing

# Using Benford's law as an epidemiological tool for COVID-19 pandemic data analysis in Greece

Kagkaras Odysseas[1], Kariofylli Aikaterini-Danai[2], Karpouzi Elisavet[3], Koufopoulou Eirini[4], Limnaios Dimitris[5], Maneta Chrysa[6]

*1. Varvakeion Model High School, Mouson 29, 15452, Athens, Greece (E-mail: odysseaskagaras@gmail.com)*
*2. Varvakeion Model High School, Mouson 29, 15452, Athens, Greece (E-mail: danaikariofylli@gmail.com)*
*3. Varvakeion Model High School, Mouson 29, 15452, Athens, Greece (E-mail: elisavetkarpouzi@gmail.com)*
*4. Varvakeion Model High School, Mouson 29, 15452, Athens, Greece (E-mail: renakouf6@gmail.com)*
*5. Varvakeion Model High School, Mouson 29, 15452, Athens, Greece (E-mail: dimL05@outlook.com.gr)*
*6. Varvakeion Model High School, Mouson 29, 15452, Athens, Greece (E-mail: chrysamaneta12@gmail.com)*

## Abstract

Have you ever imagined that nature may not be as uniform as it looks? Or that it may have a tendency toward smaller numbers? At the end of the 19th century, triggered by an observation that in logarithm tables the earlier pages (with numbers that started with 1) were much more worn than the other ones, the idea of Benford's law was born. According to the law, in collections of numbers from empirical data, the leading digit is more likely to be small. In sets that obey Benford's law, for instance, the number 1 appears as the leading digit about 30% of the time, while the number 9 less than 5%. Hence, this paper aims to examine the peculiarity of Benford's law and to apply it in the context of the critical health situation we are experiencing today, the COVID-19 pandemic, to verify the validity of case & death statistics with the criterion of whether they follow this natural anomaly. It should be mentioned that the cases that are taken at time intervals of one month may deviate from the law, while overall -when the sample size is large- they follow it at a satisfactory level. On top of that, during our research, we were faced with an illusion known to psychologists as "bias of equal probability" (Flehinger, 1966); in a nutshell, it is a human tendency to think that "real" probability implies uniformity. But does it?

*Keywords:* Benford's law, COVID-19, logarithms, epidemiology, Greece

## 1. Introduction

In the period 2020 – 2022, humanity was faced with an unprecedented health situation, the COVID-19 pandemic. In total, hundreds of millions of cases have been confirmed and millions of people have died. At the same time, successive lockdowns have shaken the psychology of citizens and radically changed our daily lives. So, in a world where everyone is exposed to misinformation, it's not easy to obtain valid information about the exact health status (cases, deaths) of our society.

In 1881, the Canadian astronomer Simon Newcomb, see [N], noticed that, in the logarithm tables, the initial pages (starting with 1) were much more worn out than the rest. The Newcomb publication is the first known example of observing a law known as Benford's law. Newcomb also suggested that the probability of a single number N being the first digit of a number is equal to log (N + 1) - log (N).

Later, in 1938, the observation of Newcomb re-emerged by Frank Benford, see [Ben], but Benford did something crucial: having recognized the statistical implications of the phenomenon, he decided to test it on a huge number of numerical data from 20 different areas. His dataset included the lengths of 335 rivers, the sizes of 3259 US populations, 104 natural constants, 1800 molecular weights, 5000 entries from a mathematical handbook, 308 numbers contained in an issue of Reader's Digest, the addresses of the first 342 people listed in American Men of Science and 418 death rates. The total number of observations used in the article was 20229.

Today, Benford's law provides for the frequency distribution of the leading digits into sets of empirical numerical data. In other words, sets that are not random, but have been created according to a physical process (such as a pandemic). Briefly, the law states that the leading digit is more likely to be small. So, in this paper we will analyse the case and death data announced by the Greek government to see if they obey this natural anomaly.

## 2. Method

The survey was implemented during the period April 2021 - February 2022 by high school students of the Mathematics Club of Varvakeion Model High School. Numerical data of cases and deaths announced by the official website of the National Public Health Organisation of Greece (https://eody.gov.gr/) for a period of 2 years (February 2020 - February 2022) were thoroughly examined. The program that was used to process the data was Microsoft Excel and the corresponding functions: *LEFT* to find the leading digit, *COUNTIF* to count the numbers that have a specific leading digit and *SUM* to find the total amount of data. Finally, suitable *frequency against initial digit diagrams* were plotted for subsequent comparison with the predictions made by Benford's law.

## 3. Benford's law mathematics

We have already mentioned that according to the law the leading digit is more likely to be small, but the exact frequency is given by a mathematical formula based on logarithmic functions:

$$P(d) = \log_{10}(d+1) - \log_{10} d = \log_{10}\left(\frac{d+1}{d}\right) = \log_{10}(1+\frac{1}{d}) \ ,$$

where $d \in \{1, \ldots, 9\}$ is the leading digit.

Therefore, the frequency of occurrence of each specific digit is given to Table 1, and Figure 1 presents another representation of the calculations.

| d | P(d) | ≅ | % |
|---|---|---|---|
| One(1) | $\log_{10}(1+\frac{1}{1})$ | 0,301 | 30,1 |
| Two(2) | $\log_{10}(1+\frac{1}{2})$ | 0,176 | 17,6 |
| Three(3) | $\log_{10}(1+\frac{1}{3})$ | 0,125 | 12,5 |
| Four(4) | $\log_{10}(1+\frac{1}{4})$ | 0,097 | 9,7 |
| Five(5) | $\log_{10}(1+\frac{1}{5})$ | 0,079 | 7,9 |
| Six(6) | $\log_{10}(1+\frac{1}{6})$ | 0,067 | 6,7 |
| Seven(7) | $\log_{10}(1+\frac{1}{7})$ | 0,058 | 5,8 |
| Eight(8) | $\log_{10}(1+\frac{1}{8})$ | 0,051 | 5,1 |
| Nine(9) | $\log_{10}(1+\frac{1}{9})$ | 0,046 | 4,6 |

*Table 1* The specific frequency of occurrence of each digit as the leading in a set of data that obeys Benford's law.

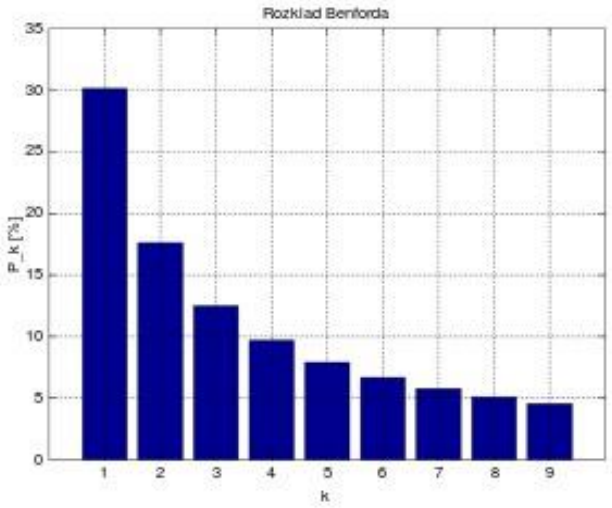


*Figure 1* The distribution of first digits according to Benford's law; each bar represents a digit, and its height is the percentage of the numbers that begin with that digit.

Thus, in sets that obey Benford's law, the number 1 appears as the leading digit in about 30% of cases, while 9 in less than 5%. On the contrary, if the digits were evenly distributed, each would appear in 11.1% of cases. Benford's law also makes provisions for the distribution of second and third digits, and possible combinations of them, see [Ber].

Today, Benford's law is used in the detection of economic frauds, in legal cases, in electoral data, in macroeconomic data, in genome data (DNA sequencing) and in the verification of the validity of scientific research.

## 4. Assumptions about the mechanism that causes the Benford distribution

Many argue that Benford's law is directly related to the fact that the universe—and therefore its content—is finite. Thus, as resources are limited, everything tends towards elimination and therefore towards smaller numbers, see [S]. Another view says that one of the reasons for the "prejudice" of nature towards the smaller numbers is that they are the building-blocks for making the big ones and therefore the small ones should be more numerous. Thus, in the context of a pandemic, first small number of cases will appear and then large ones as a consequence of the small ones, so small ones will be preceded and will probably be more numerous.

## 5. Benford: law or effect? – Possible random errors in research

In the context of mathematical science, there is a distinction between the terms "law" and "effect". On the one hand, a law is something powerful, which can be proven through mathematics and applied accurately in every case specified by the law. On the other hand, an effect is something weaker, usually experimental, and observational, without any strict proof and is based on estimations. Hence, Benford's hypothesis in 1938, due to its abstract nature, has not led to a proper proof with wide acceptance and thus seems to have more of the characteristics of the effect despite of its misleading name as a law. Nevertheless, we should note that Benford's

law does not depend on the units in which the quantities are expressed. It can be shown that this property of unit change invariance characterizes Benford's law as shown in the article by R.S. Pinkham (1961), see [P].

Therefore, for the subsequent statistical analysis of the data, we must take into consideration that “Benford's law” has the nature of an effect and not that of a law. Other random errors that will surely affect the results of our data analysis are the samples examined which are not ideal; during periods of the pandemic the number of total diagnostic tests changed, the vaccines similarly affected the number of cases, as well as the different protection measures taken by the government during each period (e.g., lockdowns).

Consequently, the pandemic during the 2 years being examined did not have a continuous course but suffered external interventions. Of particular importance is also that in the initial months of 2020, due to the more impulsive management of the situation, there is expected to be a greater influence of random errors. However, we consider that as the amount of data we examine is data of a 2-year period (724 days) and the study analyses the situation as a whole, the random errors will be minimised, increasing the precision and the accuracy of the results.

## 6. The paradox of the law, even distribution & negative control

If you observe the universe in an astrophotography, the uniformity that dominates it macroscopically will be characteristic. Thus, one would expect that even distribution is an inherent feature of nature. However, a first look into Benford's law makes the impression that nature seems to favour small numbers. One would easily think that the "real" probability implies uniformity, but once again science reminds us that nature has no obligation to be easily understood and predictable by the human species...

Of course, there are also cases where there is indeed an even distribution, such as random sets of real numbers that obey the law of Large Numbers; let us create, for example, in Excel a database of 1500 random numbers from 1 to 100000, using the *RANDBETWEEN* function. By calculating the percentage frequency of each digit as the leading digit of these numbers, we observe that it is similar for each digit (≅11%).

Moreover, the specific database of random numbers and the frequencies retrieved will be useful as a negative control for our research, that is, a set that does not have the features of the sets that obey Benford’s law (i.e., it is not made by a physical process) and hence we do not expect to show the correlation between the leading digit and the frequency predicted by the law. This will contribute to the validity of this research by showing that it is not the method of data analysis followed or some other factor that showed a *leading digit - frequency correlation* like that described by Benford, but the inherent nature of the set examined. Figure 2 shows the frequencies of leading digits of a set of random numbers compared to those provided for by Benford’s law.

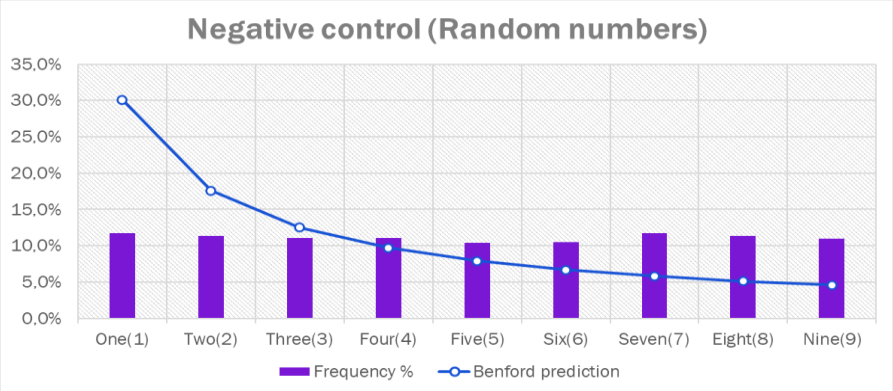


*Figure 2* The frequencies of initial digits in a set of random numbers (purple) and Benford’s prediction (blue). It is easy to observe the even distribution and the lack of correlation (horizontal line) of the variables.

# 7. A psychological point of view

We have seen that in a set of random numbers the law of Large Numbers is obeyed, that is, there is an equal probability that the leading digit will be any number. What if, though, these sets were not just random numbers, but numbers of our everyday life, created by a certain mechanism such as that of the pandemic? In such a question many would be those who would continue to believe again in a uniform distribution in their intuition.

This was examined in 1966 by B. J. Flehinger, see [F], who described it as an illusion of the human mind – also known to psychologists as the "*bias of equal probability*". It is, hence, a human tendency to consider that "real" possibility implies uniformity. However, nature may not actually be as perfect and uniform as it seems...

# 8. Results – Statistical analysis

The results of our research are presented below. Figure 3 shows the distribution of the leading digits of cases compared to that provided by Benford's law for the whole period examined (February 2020 – February 2022), while the corresponding Figure 4 is for the death data. Figure 5 and Figure 6 show data for the period of a month.

It was also examined whether the order of magnitude of the data affects the results, so Figure 9 shows distribution of death data between 1 and 10, which is more compatible with Benford's law than showing all the data regardless of their order of magnitude. Similarly, Figure 7 and Figure 8 examine cases of specific order of magnitude. The reason for examining different orders of magnitude separately is that there is periodicity; as the order of magnitude changes the leading digit becomes again 1.

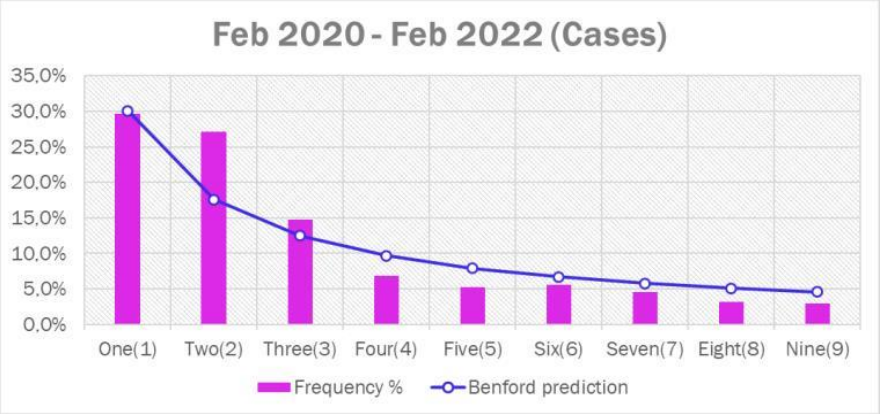


*Figure 3*

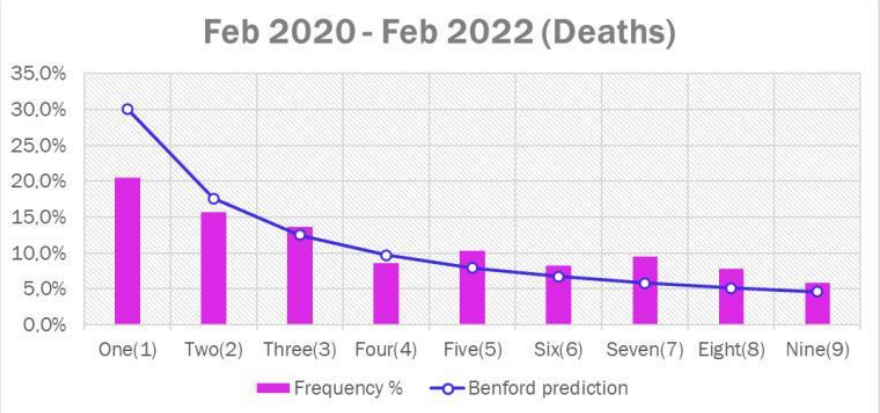


*Figure 4*

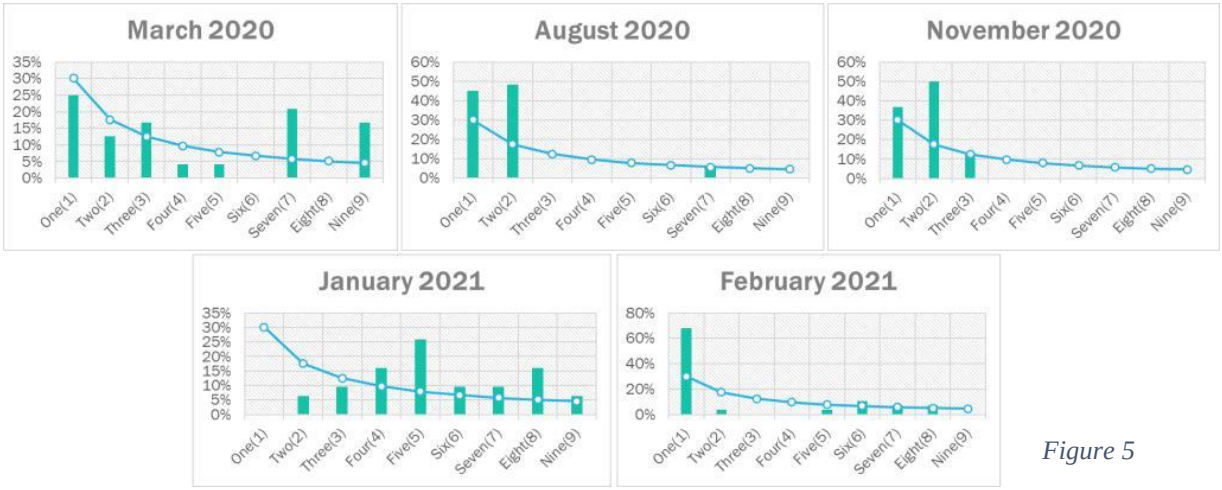


*Figure 5*

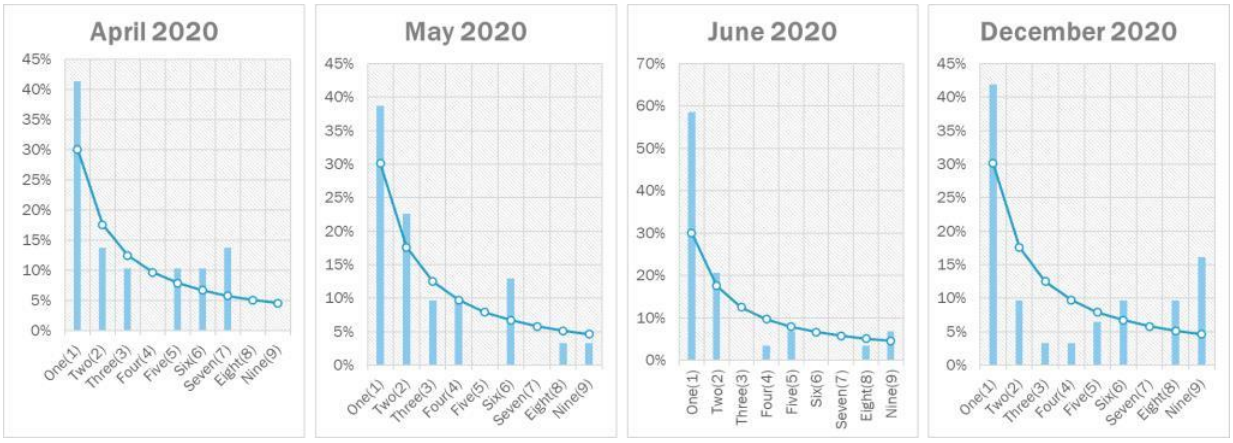


*Figure 6*

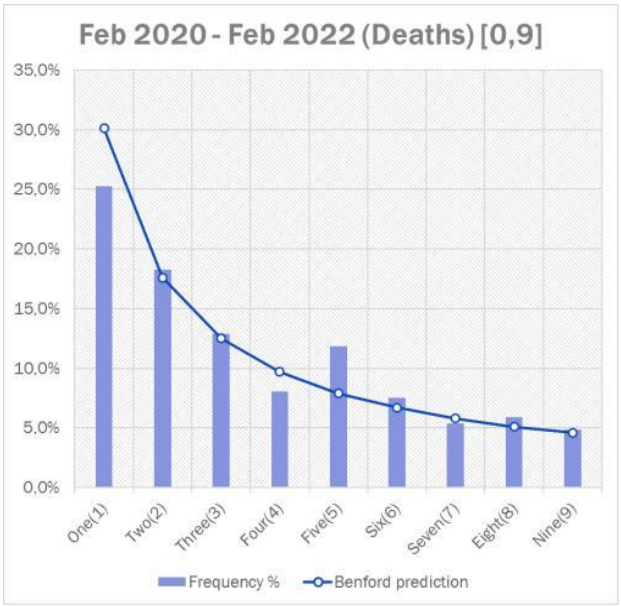


*Figure 9*


*Figure 7 & 8*

# 9. Interpreting the diagrams

After analysing the diagrams, we can conclude that as concerned the case data, throughout the period the law is largely obeyed, and it seems that our initial hypothesis is being verified. This can be attributed to the large sample size used and to the wide range of case data. Therefore, there is a *strong correlation* between initial digit and its corresponding frequency according to Benford's law. Similarly, as for the death data, again there is a pattern like Benford's, but this time there is a *weak correlation* since it is not as clear as in the previous case. This can be attributed to the smaller range of death data (from 0 to 134) compared to the range of case data (0 to 50126), at the same time the size of the data examined remains the same. As a result, a first conclusion we can make is that (due also to the logarithmic nature of Benford's law) *as the order of magnitude and the range in the numerical sets examined increases, the more obvious the effects on the unequal distribution imposed by Benford's law are.*

When it comes to diagrams for periods of a month, the patterns are sometimes more inconspicuous and sometimes more obvious. Figure 5 shows bad examples for verifying whether the law is applicable, due to the limited number and range of numerical data. Thus, there is no surplus in the smaller digits and/or there is a complete lack of data with an initial digit of a specific number. On the contrary, Figure 6 shows good examples for verifying whether the law is obeyed, as in those months there was enough cases every day, as well as a wide range of values. So, a second conclusion is that *the larger the number of data (sample size), the more accurately the law is obeyed.*

Finally, as for the diagrams that look at the whole period but focus on specific orders of magnitude, either they are more consistent with the law or not. For example, Figure 7 and Figure 8 focus on a daily number of cases from 10 to 99 and from 0 to 999 respectively, showing distinct Benford's distributions. Similarly, Figure 9 focuses on a daily number of deaths from 0 to 9, which includes 257 days out of the total of 2 years and shows a better Benford distribution than the graph (Figure 4) that looks at the total number of deaths regardless of their order of magnitude.

# 10. Conclusions

The results of the research confirm that a pandemic is a dynamic phenomenon that evolves and is characterized by continuity, i.e., each phase of the pandemic cannot be taken independently of the others. On top of that, the validity of the data announced by the government is verified since, considering random errors, the hypothesis that data of cases and deaths follow Benford's law is confirmed. Consequently, it is clear that Benford's law as a tool of statistical analysis for the science of epidemiology is very important, because, in addition to confirming or disproving data, it could also make predictions in the future about how a pandemic would evolve.

# Acknowledgement

The authors express their thanks to Prof. Lygatsikas Zenon for his support as a supervisor of this paper and his suggestions from the beginning to the end of this project.

# References


[Ben] Benford, F. (1938). *The law of anomalous numbers*. *American Philosophical Society*, Vol. 78, No. 4, pp.551-572.

[Ber] Berger, A., & Hill, T. P. (2020). *The mathematics of Benford's law: a primer*. Recovery from https://arxiv.org/pdf/1909.07527.pdf

[DK] DK. (2021, February 11). *A Treatise On Benford's Law*. Recovery from https://aenigmaenterprises.medium.com/newcomb-benford-effect-part-1-39555b4080d5

[F] Flehinger, B. (1966). *On the probability that a random integer has initial digit A. The American Mathematical Monthly*, 73:1056—1061.

[N] Newcomb, S. (1881). *Note on the frequency of use of the different digits in natural numbers*. *American Journal of Mathematics*.

[P] Pinkham, R.S. *On the distribution of first significant digits*. Ann. Math. Statist., 32:1223—1230, 1961.

[S] Sanchez, E. (2020, August 9). *Behind Benford's Law is Inherent Scarcity*. Recovery from https://ericlsanchez.medium.com/behind-benfords-law-is-inherent-scarcity-91ca9badb72e