Application of Benford's Law to Detect if COVID-19 Data is under Reported or Manipulated

Divyasha Pahuja

Bhagwan Parshuram Institute of Technology, New Delhi, India

Corresponding author: Divyasha Pahuja, Email: divyashapahuja@gmail.com

Tackling of COVID-19 has been the prime concern for all the countries because of the virus' deadly tenure across the globe. But an attempt to curb the virus requires an effective analysis of its spread. Falsified data is suspected by experts on account of policy-making, political strategies and fear handling. Benford's law, which is an effective and widely used technique to detect fraud and fabrication, is employed and applied to data made available publicly, by the authorities, in an attempt to see if the data adheres to the Benford's Law and if it is correctly reported or not. Benford's law works well on exponential data and is compared to the initial growth data of the countries, but in case of control interventions an epidemic growth model is applied to the data and then goodness-of-fit test is conducted to see if Benford's Law is satisfied for the selected three countries. Also, the level and rate of control is calculated to determine the deceleration in growth for each country.

Keywords: Benford's Law, COVID-19, Epidemic growth model, SEIR, Chi-Square test.

2021. In Rahul Srivastava & Aditya Kumar Singh Pundir (eds.), *New Frontiers in Communication and Intelligent Systems*, 85–91. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-95502-00-4-11

1 Introduction

In December 2019 the first strain of SARS-CoV-2 was found in Wuhan, Hubei province of China. Corona virus disease or COVID-19 was declared by WHO (World Health Organization) as a global pandemic and public health emergency on January 30th. Since then the novel virus has been rapidly propagating across multiple countries and threatening the lives of people globally. Research is hence required to effectively tackle the spread of COVID-19. However, data provided by the local governments and agencies is sometimes inconsistent for many reasons and may not be fully correct to analyze the situation and take measures in an appropriate manner. Therefore, an effective method to analyze the correctness of data is required. The use of Benford's Law is suggested to detect if the data is underreported or missing.

In case of no restrictions imposed, the virus would spread rapidly giving a steep rise in the number of cases and deaths. The cases would grow exponentially in a natural state and hence follow Benford's Law (BL) as any exponential distribution is known to automatically follow BL. However, in case of a controlled growth environment where a lockdown is imposed and social distancing practices are in place including practicing of personal hygiene such as wearing mask and washing of hands, as well as inoculating the people, the steep of the curve would not be of exponential degree. In such cases, if the restrictions and interventions are successful then the confirmed cases and deaths would not follow Benford's Law. For this the Benford's Law is useful for detecting if the control interventions currently being employed have been effective.

An epidemic growth model is applied accommodating the restrictions applied by countries which is then fitted to the data provided for the countries under analysis to see if the number of cases in a controlled environment follow BL.

The countries chosen are India, United States and Italy to analyze our theory because their high populations, detection of cases early on in the pandemic and rapid spread of the virus make them the perfect specimen for detecting the application of Benford's Law.

2 Benford's Law

The Benford's Law (or actually Newcomb-Benford Law, said to be first discovered by Newcomb) states that the leading digits of a number (occurring naturally) are skewed towards smaller numbers. The probability distribution of single digits in significant positions of a number are not uniform, they follow a logarithmic distribution.

Benford's Law is mainly present in all natural systems (including city populations, no. of towns in a country, count of rivers etc.) and is less likely to be found in human made systems such as identification numbers (e.g. Social security numbers, phone number, house number, car license plate number) or those which include range and maximum and minimum limits (e.g. height of a human). According to BL the probability of the first digit being 1 is 30.1%, first digit being 2 is 17.6% and so on as shown in Table 1. Table 1 show both and first and second digit Benford's distribution, as gathered from [1], [2] and [16]. The probability of first digit is calculated using the formula:

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

where d ∈ (1, . . . , 9).

Table 1. Benford's Distribution of first and second digit

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

3 Dataset

Data plays a crucial role in any analysis. So the dataset collected is important for the analysis to run with high accuracy. The dataset used for this paper is provided by John Hopkins University Center for Systems Science and Engineering (JHU CSSE). The dashboard [12] provides visual representation of the daily data for many countries from 22nd of January 2020. It includes data of confirmed cases, deaths, recoveries and vaccinations. The CSV (comma separated values) for the same can be found at [13]. The confirmed cases for each country (India, USA and Italy) are extracted from these and divided into 2 sets: before lockdown and after lockdown.

4 Methods

4.1 Exponential growth

As discussed, the countries will follow an exponential growth unless control interventions are employed. The lockdown restrictions and social distancing norms were placed at different points of time for different countries.

For India, the first case was seen on 30th of January 2020 and on 16th of March practices for social distancing were recommended [14]. But only on 25th of March a nation-wide lockdown was imposed on the citizens of the country. USA saw its first case on 21st January and "stay-at-home" orders were announced for state-wise starting from 19th of March. In Italy, the virus was detected first on 31st of January among 2 people and a lockdown was announced from 9th of March due to the sudden multiplication of virus hosts [18].

Frequencies of first digit among the daily confirmed cases during the period before lockdown is computed and compared to the Benford's first digit distribution as seen in Figure 1. The data is handled using Pandas framework and the results are represented using matplotlib library in python.

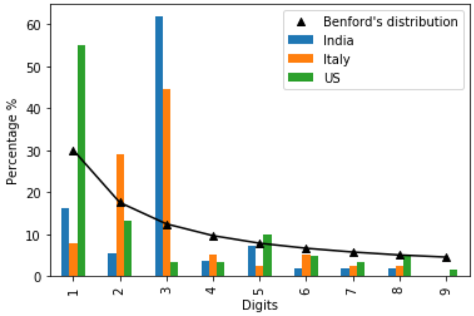

Fig. 1. First digit distribution of cases in countries before first lockdown

As we can see from Figure 1, US follow a logarithmic distribution of the first digits during the initial days without interventions, with a very high percentage for 1 as first digit. However, in case of India the digit 3 and in case of Italy digits 2 and 3 have higher frequencies and so they do not conform to Benford's law, as briefly shown in [19], even during the initial growth phase of the corona virus.

4.2 Controlled growth after restrictions

The period after first lockdown is analyzed to see if the confirmed cases follow Benford's distribution of first digits. Figure 2 is a visual representation of the same.

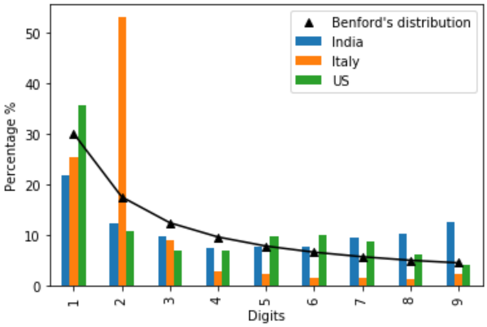

Fig. 2. First digit distribution of cases in countries after lockdown

As we can see from Figure 2, the confirmed cases in India even after lockdown do not follow Benford's distribution as higher frequencies of larger digits can be seen. Same can be said for Italy, where the frequency for digit 2 is much higher than the rest. Whereas, in US the numbers can be said to conform to the Benford's distribution even after restrictions. This could mean that the lockdown imposed was not effective. So an epidemic growth model capturing these control restrictions is fit to the COVID-19 data of these countries to determine the level of intervention.

4.3 Epidemic Growth Model

The Susceptible—Exposed—Infectious—Recovered (SEIR) which is a model built for effective analysis of how a viral disease spreads and grows is used. The SEIR is a derivative of SIR model. Many derivatives exist but SEIR is considered most appropriate for COVID-19 epidemic [20]. As can be derived from the name, the SEIR has 4 parts — susceptible individuals (S), exposed people who are not infectious (E), infectious individuals (I) and those who have recovered from the disease (R)
SEIR model consists of the following ordinary differential equations (ODEs):

$$\frac{dS}{dt} = -\frac{\beta SI}{N} \qquad (2)$$

$$\frac{dE}{dt} = \frac{\beta SI}{N} - \sigma E \qquad (3)$$

$$\frac{dI}{dt} = \sigma E - \gamma I \qquad (4)$$

$$\frac{dR}{dt} = \gamma I \qquad (5)$$

where β is the transmission rate,
γ is the recovery rate, σ is the incubation rate i.e. rate to be infected

$N = S + E + I + R$ is the whole population.

Now lockdown restrictions and social distancing measures in place change the rate of transmission, β. Therefore, to quantify the impact of these control interventions a variable ρ is included in the model. In the ODEs β would become ρβ. When ρ=1, then no lockdown restrictions have been place, if $0< \rho <1$, then control interventions have been used to stop the virus, or "flatten the curve". Lesser ρ would mean flatter curve, ρ closer to 1 would mean the restrictions in place are not severe.

Initial I is taken as the number of confirmed cases from lockdown date of the country, R is the sum of recovered people and deaths, E is calculated by multiplying 2.399 with confirmed cases 6 days from the lockdown date as suggested in [21]. β and ρ are calculated using least-square curve fitting process on the equations. It is calculated using curve_fit() function in python. The mean square error (MSE) is then computed. Then the parameters are updated using MSE, and fed to the next iteration to update the parameters as shown in [17]. This process is iterated 50 times. The initial values of β and ρ are taken as 0. The final values computed for each country are given in Table 2.

4.4 Chi-Square test

The Chi-Square Goodness-of-fit Test is used to analyze the statistical goodness of fitting of the data i.e. the gap between the expected (the Benford distribution) and the observed data. The chi-square for our cause can be calculated by:

$$\chi^2 = \sum_{i=0}^{9} \frac{(o_i - e_i)^2}{e_i} \qquad (6)$$

where c is the degree of freedom, o_i is the observed frequency, e_i is the expected frequency. According to [15], the objective is to test a hypothesis and see if the variable follows the theoretical distribution i.e. the null hypothesis. In our case null hypothesis (H_o) and alternative hypothesis (H_A) are as follows:

H_o : First significant digit follows Benfords Law
H_A : First significant digit does not follow Benfords Law

So for 5% significance level (p-value), the critical value would be χ^2 (8) = 15.51.
χ^2 value of our data less than the critical value would be accepted in null hypothesis and greater than this would be rejected and belong to the alternative hypothesis.

The goodness of fit with the JHU COVID-19 data is discussed in the result section.

5 Result and Discussion

For before lockdown, when the countries are expected to have an exponential growth, it is seen that only US seems to follow the Benford's Law. The discrepancies seen in the other 2 countries, as seen in Figure 1, could be attributed to the fact that the countries were not testing many people initially or did not have the capacity to test the suspected people as it can be seen in India that the cases reported were stuck at 3 for about a month (in the data provided by JHU), as is also mentioned in [18] in case of Italy, or simply the scenario that the cases were underreported or misreported. After lockdown announcement and social distancing measures were put in place, the epidemic growth model is used to calculate the growth control parameter (ρ) for each country and see how effective the restrictions have been and χ^2 calculated to see if the numbers follow the Benford's distribution, as shown in Table 2.

Table 2. SEIR model in presence of restrictions and χ^2 for Benford's Law

Country	β (Growth rate)	ρ (Control rate)	χ^2	p-value
India	0.782	0.75	4.59	0.72
US	0.476	0.91	1.29	0.99
Italy	1.021	0.69	3.56	0.855

From Table 2 it can be concluded that US indubitably conforms to the null hypothesis that the confirmed cases of the country follow the Benford's distribution. Hence, the lockdown restrictions imposed were not effective (with control parameter of 0.91). For Italy and India the null hypothesis cannot be rejected, but have slightly better virus deceleration parameters of 0.69 and 0.75 respectively. So they have managed to curb the virus at a much better rate.

6 Conclusion and Future Work

In this work, Benford's law is used to analyze and detect if the the reported cases by the government or respective authorities of the country are correct or not. It was found that discrepancies were found in the initial phase of exponential growth for countries India and Italy. It could be attributed to no testing kits available. An SEIR epidemic growth model is adopted to analyze the numbers and situations after lockdown restrictions were laid, the countries are found to to somewhat conform to Benford's distribution. The level of control and restrictions is also computed and analyzed for each country.

This research work has a vast scope for future work. The current work only employs a static SEIR model, it works on assumption that the population will remain constant considered. A new model that takes in vital dynamic values and parameters is to be followed in the future work. And the research can be expanded to more countries and also include deaths for better results. As more and more people are vaccinated and herd immunity is achieved the dynamics of the research will change.

References

[1] F. Li et al., "Application of Benford's Law in Data Analysis", *J. Physics: Conf. Series,* 1168, 2019.

[2] S. W. Lanham, "Analyzing Big Data with Benford's Law: A Lesson for the Classroom", *American J. Business Edu.,* vol. 12, no. 2, 2019.

[3] M. J. Nigrini, "A Taxpayer Compliance Application of Benford's Law", *The J. the American Taxation Assoc.* vol. 18, pp. 72–91, 1996.

[4] "Fraud Detection using Benford's", https://towardsdatascience.com/frawd-detection-using-benfords-law-python-code-9db8db474cf8, last accessed 2021/05/21.

[5] I. Aamo and S. Caleb, "On the Use of Benford's Law to Detect JPEG Biometric Data Tampering", *J. Inform. Security,* vol. 8, pp. 240–256, 2017.

[6] D. Fu, Y. Q. Shi and W. F. Su, "A generalized Benford's Law for JPEG Coefficients and its Applications in Image Forensics", in *Proceedings of SPIE - The International Society for Optical Engineering*, 2007.

[7] J. Golbeck, "Benford's Law can Detect Malicious Social Bots", *First Monday,* vol. 24, 2019.

[8] D. Gammerman and F. L. Antunes, "Statistical Analysis of Brazilian Electoral Campaigns via Benford's Law", *Physica A: Stat. Mech. its Appl.,* vol. 496, pp. 171–188, 2018.

[9] K. Y. Wook, J. Han C. S. Rak, "Detection of Possible Match-fixing in Tennis Games", in *Proc. Of the 6th Int. Cong. on Sport Sci. Res. and Tech. Support,* 2018, pp. 124-131.

[10] M. A. Amouzegar, K. Moshirvaziri and D. Snyder, "BENFORD'S LAW AND ITS APPLICATION TO MODERN INFORMATION SECURITY", 2018.

[11] L. Sun et al., “Detection and Classification of Malicious Patterns In Network Traffic Using Benford’s Law”, *Asia-Pacific Signal and Inform. Proc. Assoc. and Annual Summit and Conf.*, pp. 864-872, 2017.

[12] “JHU Coronovirus Resource Centre”, https://coronavirus.jhu.edu/, last accessed 2021/08/09.

[13] “Novel Coronavirus Cases Data”, https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases, last accessed 2021/08/09.

[14] A. Ghosh, S. Nundy and T. K. Mallick, “How India is dealing with COVID-19 pandemic”, *Sensors Int.,* vol. 1, 100021, 2020.

[15] E. Gijka, L. Basha and L. Puka, “An Analysis of the Reliability of Reported COVID-19 Data in Western Balkan Countries. Advances in Science”, *Tech. and Eng. Systems J.,* vol. 6, no. 2, pp. 1055-1064, 2021.

[16] J. F. Coeurjolly, “Digit analysis for Covid-19 reported data”, arXiv:2005.05009(1) (2020)

[17] K. B. Lee, S. Han and Y. Jeong, “COVID-19, flattening the curve and Benford’s law”, *Physica A: Stat. Mech. and its Appl.,* vol. 559, 2020.

[18] C. Koch and K. Okamura, “Benford’s Law and COVID-19 reporting”, *Eco. Lett.,* vol. 196, 2020.

[19] V. H. Moreau, “Inconsistencies in countries COVID-19 data revealed by Benford’s Law”, *Model Assisted Stat. and Appl.*, vol. 16, pp. 73-79, 2021.

[20] “Statistical Modeling for the Prediction of Infectious Disease”, https://www.frontiersin.org/articles/10.3389/fpubh.2021.645405/full.

[21] F. A. B. Hamzah et al., “Corona Tracker: world-wide COVID-19 outbreak data analysis and prediction”, *Corona Tracker,* 2020.