Lies, damned lies, and statistics: The uncertainty over COVID-19 numbers in India

Kiran Mahasuar

Strategic Management Area, Indian Institute of Management, Kozhikode, India

Correspondence
Kiran Mahasuar, Room No: A-19, Indian Institute of Management, Kozhikode, Kerala 673570, India.
Email: kiranj13fpm@iimk.ac.in

This paper intends to ascertain the veracity of reported data on deaths and testing pertaining to the novel coronavirus in India. We use a widely used forensic audit technique called Benford’s law to analyze the data, and our findings suggest anomalies in the reported numbers and the reported data for most of the states do not adhere to the Benford distribution. The implications of these findings are manifold, especially on the trajectory of policy-making, vaccination strategy, and preparedness for future waves and new variants. We strongly argue for the need for a robust data collection and reporting mechanism, creating a central data repository, and instituting a data-driven policy framework as key steps in the process management bulwark for managing such future pandemics and other events concerning public health.

1 | INTRODUCTION

The outbreak and spread of the novel coronavirus (COVID-19) has been catastrophic for India and has impacted the people and the economy severely. The policy response to the pandemic has been largely inadequate and muted and consequently, the second-most populous country is staring at an unprecedented humanitarian crisis with the healthcare system on the brink of collapse. Experts cite the paucity of accurate data and sampling as a major constraint in envisaging the uneven pattern of the spread, genome sequencing in cluster level investigations, identifying new variants, and predictive modeling of future waves in India (Ethiraj, 2021; Mallapaty, 2021). The reported data have also been under the scanner with anecdotal evidence suggesting massive under-reporting of deaths in particular. This issue has also attracted considerable negative press, including in premier outlets like Time Magazine and New York Times (Bajekal, 2021; Gettleman et al., 2021). To be fair such concerns were also raised for the reported data from China, but Koch and Okamura (2020) debunk that speculation and find no evidence that the Chinese authorities massaged the COVID-19 statistics.

The aforementioned concerns and the issues of data integrity engendered our investigation where we focus on decoding the COVID-19 data pertaining to deaths and tests as reported from India. We use Benford’s law—a widely used and acclaimed technique in forensic audit—to assess the veracity of the reported data on Covid deaths and tests. This empirical research assumes a lot of significance in the context of high levels of distrust and skepticism regarding the Covid numbers and ensuing policy responses from the government. The paper is structured as follows: In the second section, we discuss the methods and materials used for this study. In the third section, we discuss the findings threadbare, in the fourth section, we discuss the ramifications of the anomaly of underreporting and policy implications and in the last section, we conclude the study along with limitations and future research directions.

2 | METHODS AND MATERIALS

2.1 | Brief introduction of Benford’s law

The origin of Benford’s law (also known as first-digit law or Newcomb-Benford law) can be traced to Newcomb’s work published in the American Journal of Mathematics (Newcomb, 1881). The mathematician observed that the logarithms book at the library was more worn on the front pages and less worn on the back pages. Newcomb subsequently devised a formula for calculating the probability of any non-zero initial digit of a number. Benford (1938) revisited this pattern several years later and confirmed its consistency in a myriad of distributions like home addresses, river lengths, and so on. The incidence of the first significant digits (FSD), according to Benford, follows a logarithmic distribution:
where $P_k$ is the probability for a given number of having $k$ as the FSD in a distribution. Thus, the probability of occurrence of each digit is as follows (see Table 1).

Benford’s law is scale-invariant and irrespective of changes in the unit of measurement in the data, the compliance does not get affected (Mir et al., 2014). Given the empirical evidence of the higher occurrence of lower digits vis-à-vis higher digits, a dataset derived organically should ideally follow Benford’s theoretical distribution. An anomalous result demonstrating non-compliance with this law indicates a possible data manipulation, and therefore the dataset and its sources must be thoroughly investigated. Benford’s law has been extensively used to detect frauds in economic, behavioral, and accounting data (Varian, 1972). Early applications in accounting include work by Carslaw (1988) and Thomas (1989), Nigrini (1996, 2005) used Benford’s law to examine the phenomena of tax evasion and the Enron fraud. Other notable works in this space include works by Johnson (2005), who analyzed the earnings per share of various industries, and Henselmann et al. (2013), who analyzed financial reports of US companies. Benford’s law has also been used in governance and policy-making. Notable instances include evaluating if economic data has been manipulated in a distribution. Thus, the probability of occurrence of each digit is as follows (see Table 1).

### 2.2 Formal tests

We have used Pearson’s Chi-square test (Pearson, 1900) as a significance test to ascertain if the “true” distribution follows the theoretical (Benford) distribution and if the sample comes from a distribution with a certain probability density function (González, 2020). The Chi-square statistic is calculated as follows:

$$ P_k = \log_{10} \left( 1 + \frac{1}{k} \right) \text{contd...} = 1,2,...,9 \quad (1) $$

In addition, we have supplemented this analysis with another goodness of fit test, that is, Kolmogorov–Smirnov (KS) test wherein the empirical distribution function $F^e$ for $n$ independent and identically distributed ordered observations $x_i$ is defined as:

$$ f^e(x) = \frac{1}{n} \sum_{j=1}^{n} I_{(-\infty,x]} X_j $$

where $I_{(-\infty,x]} X_j$ is the indicator function, equal to 1 if $X_j \leq x$ and equal to 0 otherwise.

The KS statistic for a given cumulative distributive function $F^x$ is

$$ D_n = \sup_x |f^e(x) - F(x)| \quad (4) $$

where $\sup_x$ is the supremum of the set of distances. In general, the statistic takes the largest absolute difference between the two distribution functions across all $x$ values.

Both the Chi-square test and KS test can present type I error in large samples, and therefore their compliance with Benford’s law is circumspect in scenarios where we reduce the difference in the proportions (Barney & Schulzke, 2016; Druic et al., 2018; González, 2020). To alleviate this concern, we use the mean average deviation to check the concordance of the frequency distribution with Benford’s law.

$$ \text{MAD} = \frac{1}{N} \sum_{i=1}^{n} |p_i(o) - p_i(e)| \quad (5) $$

where $p_i(o)$ is the proportion of observations observed for class $i$ and $p_i(e)$ is the expected proportion for class $i$ according to Benford’s law. The adjusted mean absolute deviation (MAD) critical value ranges developed by Drake and Nigrini (2000) are shown in Table 2. We have used the same to define the conformity levels of our data.

### 2.3 Sources of data

The data for this study was sourced from the open-access Covid database maintained in Kaggle, which is a repository of data pertaining to COVID-19 testing, deaths, and vaccination sourced from the Ministry of Health and Family Welfare of the government of India (Raj Kumar & Devakumar, 2021). We have looked at the daily reported deaths for

### Table 1 Expected digit frequencies based on Benford’s law

| Digit | First position | Second position |
|-------|----------------|-----------------|
| 0     | n/a            | 0.11968         |
| 1     | 0.30103        | 0.11389         |
| 2     | 0.17609        | 0.19882         |
| 3     | 0.12494        | 0.10433         |
| 4     | 0.09691        | 0.10031         |
| 5     | 0.07918        | 0.09668         |
| 6     | 0.06695        | 0.09337         |
| 7     | 0.05799        | 0.09035         |
| 8     | 0.05115        | 0.08757         |
| 9     | 0.04576        | 0.085           |

### Table 2 Mean absolute deviation critical value ranges

| Close conformity | 0.00–0.006 | 0.00–0.008 |
| Acceptable conformity | 0.006–0.012 | 0.008–0.010 |
| Marginal conformity | 0.012–0.015 | 0.010–0.012 |
| Non-conformity | >0.015 | >0.012 |
the top 15 states (in terms of population) from April 1, 2020 until May 19, 2021. As a secondary measure of robustness check, we have also looked at the reported deaths data for the second wave exclusively for 10 states with the highest reported deaths. To supplement our analysis, we have also looked at the reported testing data of the aforementioned 15 states from April 1, 2020 till May 19, 2021. The underlying premise is that if the sampling methodology remains unchanged, the number of verified cases and deaths will follow an exponential trend and obey Benford's law.

3 | RESULTS

3.1 | Analysis of reported deaths

The results for the analysis of the reported Covid deaths in the selected period are presented below in Table 3. The 15 states represented in this table constitute 82% of India’s total population, and the reported deaths in 12 out of these 15 states do not correspond to the Benford distribution. These 12 states constitute approximately 65% of India’s population. Only the reported death numbers in Bihar, Karnataka, and Andhra Pradesh correspond to the Benford distribution based on the Chi-square statistic. However, as noted in previous works on Benford’s Law, the Chi-Square test is sensitive to small differences and this impacts the statistical significance. Hence, we supplement the results with the KS test. The supremum and p-values for the KS test statistic also confirm the possibility of manipulated data for several states. Further, we rely on the MAD statistics to corroborate our findings and check the levels of conformity with Benford’s law. We find that 11 out of 15 states show non-conformity as per critical value ranges cited in Table 2. Bihar shows close conformity (MAD = 0.005), Karnataka shows marginal conformity (MAD = 0.013), and Andhra Pradesh shows acceptable conformity (MAD = 0.007).

We also repeated a similar exercise for the reported deaths for the second wave exclusively, which began in 2021 (February 2, 2021

| State          | Population (in crores) | Reported deaths | Chi-Sq  | p-value | Supremum (KS) | p-value | MAD    | Conformity         |
|----------------|------------------------|-----------------|---------|---------|---------------|---------|--------|-------------------|
| Delhi          | 2                      | 22,111          | 30.03   | 0.0002  | 5.74%         | 0.2733  | 0.018  | Non conformity    |
| Maharashtra    | 12.2                   | 83,777          | 22      | 0.0049  | 8.33%         | 0.0545  | 0.017  | Non conformity    |
| Chhattisgarh   | 2.9                    | 12,036          | 26.6    | 0.0008  | 13.10%        | 0.0042  | 0.019  | Non conformity    |
| Karnataka      | 6.6                    | 22,838          | 12.1    | 0.1467  | 4.45%         | 0.469   | 0.013  | Marginal          |
| Tamilnadu      | 7.6                    | 18,369          | 66.2    | 0       | 11.73%        | 0.004   | 0.022  | Non conformity    |
| Haryana        | 2.9                    | 6,923           | 29.76   | 0.0002  | 10.05%        | 0.0298  | 0.021  | Non conformity    |
| Kerala         | 3.5                    | 6,612           | 203.58  | 0       | 34.57%        | 0       | 0.043  | Non conformity    |
| Andhra Pradesh | 5.2                    | 9,580           | 9.68    | 0.288   | 5.79%         | 0.3077  | 0.007  | Acceptable        |
| West Bengal    | 9.7                    | 13,576          | 166.22  | 0       | 16.73%        | 0       | 0.034  | Non conformity    |
| Gujarat        | 6.8                    | 9,269           | 75.96   | 0       | 20.28%        | 0       | 0.029  | Non conformity    |
| Rajasthan      | 7.7                    | 7,080           | 125.91  | 0       | 26.55%        | 0       | 0.056  | Non conformity    |
| Madhya Pradesh | 8.2                    | 7,139           | 64.88   | 0       | 10.44%        | 0.0126  | 0.013  | Marginal          |
| Uttar Pradesh  | 22.5                   | 18,072          | 24.98   | 0.0016  | 4.71%         | 0.424   | 0.017  | Non conformity    |
| Odisha         | 4.4                    | 2,357           | 112.95  | 0       | 28.32%        | 0       | 0.056  | Non conformity    |
| Bihar          | 12                     | 4,039           | 8.92    | 0.349   | 6.34%         | 0.2461  | 0.005  | Close             |

| State          | Population (in crores) | Reported deaths | Chi-Sq  | p-value | Supremum (KS) | p-value | MAD    | Conformity         |
|----------------|------------------------|-----------------|---------|---------|---------------|---------|--------|-------------------|
| Maharashtra    | 12.2                   | 32,695          | 41.717  | 0       | 0.222         | 0.005   | 0.038  | Non conformity    |
| Delhi          | 2                      | 11,258          | 48.869  | 0       | 0.19          | 0.026   | 0.036  | Non conformity    |
| Karnataka      | 6.6                    | 10,621          | 21.319  | 0.006   | 0.137         | 0.135   | 0.03   | Non conformity    |
| Uttar Pradesh  | 22.5                   | 9,414           | 26.305  | 0.001   | 0.177         | 0.047   | 0.027  | Non conformity    |
| Chhattisgarh   | 2.9                    | 8,335           | 5.27    | 0.728   | 0.07          | 0.594   | 0.012  | Acceptable        |
| Tamilnadu      | 7.6                    | 6,013           | 11.09   | 0.197   | 0.111         | 0.268   | 0.01   | Acceptable        |
| Gujarat        | 6.8                    | 4,882           | 49.99   | 0       | 0.314         | 0       | 0.062  | Non conformity    |
| Haryana        | 2.9                    | 3,905           | 15.406  | 0.052   | 0.166         | 0.078   | 0.035  | Non conformity    |
| West Bengal    | 9.7                    | 3,403           | 14.043  | 0.081   | 0.094         | 0.393   | 0.023  | Non conformity    |
| Kerala         | 3.5                    | 2,869           | 61.584  | 0       | 0.331         | 0       | 0.064  | Non conformity    |
till May 19, 2021). We chose 10 states based on the reported death numbers in the second wave. Cumulatively, these states saw close to 94,000 deaths in the second wave and amount to almost 30% of the total deaths reported in India across all states and union territories. The results are presented in Table 4.

Based on MAD statistics for this specific data pertaining to the second wave, the reported deaths across most states (barring Tamilnadu and Chhattisgarh) do not conform with Benford’s law. Based on Chi-square and KS tests, the reported deaths in West Bengal, Haryana, Tamilnadu, and Chhattisgarh seem to follow Benford’s law. We present the actual versus expected probability of the initial digit occurrence in accordance with Benford’s law for a few selected states in Figure 1.

Based on our analysis for the cumulative data and the second wave-specific data, we conclude that the possibility of data manipulation in reported deaths for several Indian states cannot be ruled out. Therefore, the data necessitates further examination at the district and town levels to identify the actual sources of malfeasance.

### 3.2 Analysis of reported testing numbers

The reported data on deaths is closely interlinked to the state-level tests’ robustness and subsequent mitigation measures. Therefore, we also examined the conformability of Benford’s law to the testing

![FIGURE 1 Actual versus Expected Probabilities of Occurrence of First Digit for Select States (Death Data)](wileyonlinelibrary.com)
numbers reported by the states. The results are reported in Tables 5 and 6. Unsurprisingly, we found a significant deviation in the reported values vis-à-vis the expected values. In the first digit test, we found the reported testing numbers of all the 15 states not conforming to the theoretical distribution as per Benford’s law. We also undertook the second digit test for all the 15 states to see if the observed distribution yields a better approximation of the Benford predicted distribution. The findings suggest that the reported testing numbers for Uttar Pradesh, Bihar, Madhya Pradesh, and Tamilnadu merit further investigation as they fail both the first and the second digit tests ($p$-values < 0.05 and MAD values > critical range). The reported testing numbers for Delhi, Gujarat, and West Bengal deserve further scrutinizing as the Chi-square statistic for both the first and second digits does not conform to Benford’s law. Delhi has been under the scanner for falling test rates in the second phase due to lab technicians’ shortage (Dasgupta, 2021). Similarly, there have been concerns about

### Table 5: Analysis of reported testing numbers (first digit)

| State          | Population (in crores) | Reported tests | First digit | Chi-Sq | $p$-value | Supremum (KS) | $p$-value | MAD | Conformity   |
|----------------|------------------------|----------------|-------------|--------|-----------|---------------|-----------|-----|--------------|
| Uttar Pradesh  | 22.5                   | 45,225,835     | 259.78      | 0      | 36.31%    | 0             | 0.063     | Non conformity |
| Maharashtra    | 12.2                   | 31,572,709     | 134.29      | 0      | 24.15%    | 0             | 0.029     | Non conformity |
| Bihar          | 12                     | 28,326,924     | 124.18      | 0      | 17.92%    | 0             | 0.047     | Non conformity |
| West Bengal    | 9.7                    | 11,567,340     | 297.26      | 0      | 12.46%    | 0.0018        | 0.05      | Non conformity |
| Madhya Pradesh | 8.2                    | 8,917,174      | 167.67      | 0      | 20.29%    | 0             | 0.042     | Non conformity |
| Rajasthan      | 7.7                    | 9,956,336      | 76.19       | 0      | 17.32%    | 0             | 0.025     | Non conformity |
| Tamilnadu      | 7.6                    | 25,430,272     | 418.44      | 0      | 40.83%    | 0             | 0.034     | Non conformity |
| Gujarat        | 6.8                    | 20,462,877     | 267.27      | 0      | 24.53%    | 0             | 0.049     | Non conformity |
| Karnataka      | 6.6                    | 28,065,593     | 115.85      | 0      | 15.01%    | 0.0001        | 0.03      | Non conformity |
| Andhra Pradesh | 5.2                    | 18,138,507     | 172.69      | 0      | 26.76%    | 0             | 0.045     | Non conformity |
| Odisha         | 4.4                    | 10,996,481     | 268.05      | 0      | 21.24%    | 0             | 0.048     | Non conformity |
| Kerala         | 3.5                    | 18,141,430     | 257.24      | 0      | 23.80%    | 0             | 0.049     | Non conformity |
| Chhattisgarh   | 2.9                    | 8,305,206      | 104.79      | 0      | 11.42%    | 0.0057        | 0.043     | Non conformity |
| Haryana        | 2.9                    | 8,408,489      | 125.53      | 0      | 17.68%    | 0             | 0.032     | Non conformity |
| Delhi          | 2                      | 18,408,865     | 272.11      | 0      | 33.56%    | 0             | 0.039     | Non conformity |

### Table 6: Analysis of reported testing numbers (second digit)

| State          | Population (in crores) | Reported tests | Second digit | Chi-Sq | $p$-value | Supremum (KS) | $p$-value | MAD | Conformity for both digits | Level of conformity for both digits |
|----------------|------------------------|----------------|--------------|--------|-----------|---------------|-----------|-----|---------------------------|-------------------------------------|
| Uttar Pradesh  | 22.5                   | 45,225,835     | 40.16        | 0      | 0.12      | 0.0046        | 0.0146    |     | Non conformity            | Both digits                          |
| Maharashtra    | 12.2                   | 31,572,709     | 10.26        | 0.33   | 0.06      | 0.2459        | 0.0057    |     | Close                     | Second digit only                    |
| Bihar          | 12                     | 28,326,924     | 38.1         | 0      | 0.15      | 0.0002        | 0.0206    |     | Non conformity            | Both digits                          |
| West Bengal    | 9.7                    | 11,567,340     | 38.78        | 0      | 0.09      | 0.0271        | 0.0096    |     | Acceptable                | Second digit only                    |
| Madhya Pradesh | 8.2                    | 8,917,174      | 19.73        | 0.02   | 0.1       | 0.0159        | 0.0151    |     | Non conformity            | Both digits                          |
| Rajasthan      | 7.7                    | 9,956,336      | 11.69        | 0.231  | 0.05      | 0.3472        | 0.0091    |     | Acceptable                | Second digit only                    |
| Tamilnadu      | 7.6                    | 25,430,272     | 60.26        | 0      | 0.16      | 0             | 0.0254    |     | Non conformity            | Both digits                          |
| Gujarat        | 6.8                    | 20,462,877     | 22.29        | 0.008  | 0.08      | 0.0787        | 0.0109    |     | Marginal                  | Second digit only                    |
| Karnataka      | 6.6                    | 28,065,593     | 9.5          | 0.393  | 0.05      | 0.387         | 0.0113    |     | Marginal                  | Second digit only                    |
| Andhra Pradesh | 5.2                    | 18,138,507     | 12.02        | 0.212  | 0.05      | 0.3565        | 0.0079    |     | Close                     | Second digit only                    |
| Odisha         | 4.4                    | 18,141,430     | 7.14         | 0.623  | 0.04      | 0.4734        | 0.0099    |     | Acceptable                | Second digit only                    |
| Kerala         | 3.5                    | 8,305,206      | 1.83         | 0.994  | 0.03      | 0.7024        | 0.0035    | Close|                            | Second digit only                    |
| Chhattisgarh   | 2.9                    | 8,408,489      | 8.99         | 0.438  | 0.03      | 0.7163        | 0.0093    |     | Acceptable                | Second digit only                    |
| Haryana        | 2.9                    | 18,408,865     | 5.42         | 0.796  | 0.04      | 0.4992        | 0.0033    |     | Close                     | Second digit only                    |
Covid data massaging in Gujarat (Desai, 2020) and the impact of state elections on slowing the pace of Covid testing in West Bengal (Sharma, 2021). The adverse reportage regarding Covid testing in the popular press suggests that the anomalous results and non-conformance with Benford’s law are not a mere aberration and indicate a systemic failure in these states.

4 | DISCUSSION

The incorrect reporting of testing and death numbers has several ramifications for policy interventions pertaining to public health. Firstly, it defeats the purpose of genome sequencing and thus delays the identification of super-spreaders or hotspots. Understanding the viral genomic sequence is key to developing treatments and vaccines that target certain functions of the virus and prepare for future changes as the virus evolves and mutates. There is considerable evidence regarding the lack of acracy (at the government level) in appreciating the importance of genome sequencing. We posit that the genesis of this under-preparedness and gross underestimation of the impact of the second wave, in particular, can be traced to massive underreporting of death numbers and the spread of infection. India’s rank is abysmal in genetic sequencing of coronavirus and data sharing with other countries. According to Global Initiative on Sharing Avian Influenza Data, India has shared only 6,200 genome sequences between January 2020 and April 2021 (Pulla, 2021; Ramakrishnan, 2021). The nodal body for genome sequencing in India, the Indian SARS CoV-2 Consortium on Genomics, was formed in December 2020 with the primary objective of analyzing 5% of positive samples from each state and 100% of samples from the international tourists who test positive. The progress has been largely tepid, with only approximately 15,000 samples sequenced by April 2021. INSACOG is plagued with teething issues like funding constraints, regulatory restrictions in importing key raw materials, data access, and lackadaisical approach of various states in ensuring the transport of samples to the labs. This bureaucratic red tape led to a considerable delay in identifying the B.1.617 variant, which wreaked havoc in Maharashtra (Acharjee, 2021; Pulla, 2021).

Secondly, in the absence of strict enforcement of Indian council of medical research (ICMR) guidelines on delineating COVID-19 deaths from the other deaths occurring in the hospital, there is no parity in the manner in which COVID-19 deaths are construed and counted. There are reported instances of hospitals omitting deaths due to comorbidities from the Covid deaths tally (Bhattacharya, 1938). Thirdly, since the reported numbers suggested that the pandemic was under control, some states chose to lower their guard. They chose to dismantle the temporary infrastructure like public spaces converted to special Covid centers. Furthermore, they also delayed the lockdown during the onset of the second wave leading to the subsequent uncontrolable crisis.

Lastly, the catastrophic effect of COVID-19 and the consequent breakdown of healthcare systems and infrastructure direct toward a dire need for forward-looking collaborative strategies. As governments, firms, and individual citizens attempt to enhance the responsiveness of the battered healthcare system, the role of the robust supply chain for essentials, deploying appropriate tools and technology for diagnosis, dynamic and agile response strategies, and synchronized policy-making at state and central level assumes much importance. Given the challenges associated with the issue, it is important to understand which policy-making may evolve and possible solutions being looked at by the state and central governments. This issue is challenging as healthcare as a subject is incredibly complex. Although health is a state subject, the central government is the keyactor in designing policies and implementing large-scale public health projects like the pulse polio immunization program. Rolling out a successful public health initiative necessitates the involvement of stakeholders from multiple domains (which includes both private, government, and non-government players) connected to the ecosystem. A few suggested measures in this context for the policymakers would be to engage the junior MBBS, and the ANM diploma students attend to Covid patients, making the relevant datasets available to epidemiologists and public health professionals, a more interactive usage of the Arogya Setu app and develop a standardized operating procedure for undertaking serosurveys and leveraging that data to predict infection fatality rates in clusters.

5 | CONCLUSION, LIMITATIONS, AND FUTURE RESEARCH DIRECTIONS

The anomalies in the data and the consequent non-conformance to Benford’s law underscore the absence of standardized surveillance, data collection, and reporting. Moreover, this issue is further compounded by the lack of a centralized data repository for the use of data scientists, public policy experts, and epidemiologists. This underlying deficiency also means that most states and the central government are plagued by a lack of real-time visibility beyond major cities. This drawback poses major challenges to epidemiologists and public health professionals from predicting the spread of the pandemic to other clusters, formulating the strategy for designing containment zones, and modeling the future waves (Ethiraj, 2021). This handicap was evident when the system failed to envisage the gargantuan impact of the second wave, and the extent healthcare infrastructure across the states failed to cope with the multifold increase in cases. The possibility of under-reported death numbers and testing data is an outcome of this malaise, and it has three major repercussions as follows. a) lack of understanding regarding the actual path traversed by the virus across the states, b) the vulnerabilities for the future waves, and c) what should be our ideal vaccination strategy.

It must be noted that the distribution not adhering to Benford’s law merely points out that the data is anomalous, and the non-compliance is not conclusive evidence of data manipulation, and further scrutinizing is necessary to draw such a strong conclusion. Our contribution in this regard is limited to detecting the presence of anomalous data in the reported numbers. Secondly, another limitation is that the data collected from the various states represent aggregated data rather than district or town level data. Although the large data
distributions embedded in an aggregated number will enhance the
effectiveness of Benford’s law to identify aberrations, the district or
town level data is more helpful in identifying specific clusters where
erroneous reporting or manipulation might have occurred.

Our analysis is primarily concerned with the composite data from
the onset of the pandemic till May 19, 2021, and we have not looked at
the impact of lockdowns and shutdowns imposed by the Govern-
ments from time to time. In line with Koch and Okamura (2020), we
expect the pre-lockdown period to follow the Benford law and the
period post-lockdown to be disruptive and not adhere to the
Benford distribution. Future research could look at this pre-
lockdown and post-lockdown period-specific analysis to validate this
hypothesis. Another avenue for further research is a comparative
analysis of excess deaths based on the data on death certification,
which can be sourced from the local government bodies like the
municipal corporations. Preliminary studies in this direction show a
multifold increase in deaths in 2021 vis-à-vis 2019 and 2020
(Ramani & Radhakrishnan, 2021; Tumbe, 2021).

We strongly advocate for the need for a robust data collection
mechanism and instituting a data-driven policy-making framework at
the central and state levels. We expect this paper to foster further research and debates regarding the fallibility of extant data in predicting outcomes and provide an impetus to data-driven policy-
making.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are available from
the corresponding author upon reasonable request.

ORCID
Kiran Mahasuar  https://orcid.org/0000-0002-3849-8479

ENDNOTES
1 India’s population is estimated to be 139.1 billion according to https://
www.worldometers.info/world-population/india-population/ (Last
accessed on 28th May 2021)
2 Genome sequencing is a technique that reads and interprets genetic
information found within DNA or RNA. Read more at https://www.who.
int/publications-detail-redirect/9789240018440
3 The Pulse Polio Immunization Programme was rolled out in India on
October 2, 1994, when India accounted for around 60% of the global
polio cases. The last polio case in India was reported a decade ago in
Howrah on January 13, 2011, and the country has been free of polio
(Source: https://www.who.int/india/news/feature-stories/detail/110-
million-children-vaccinated-in-the-country-s-first-polio-drive-of-the-
dercade, last accessed on May 29, 2021)
4 Bachelor of Medicine and Bachelor of Surgery (MBBS) is a professional
undergraduate degree in medical science in India. The duration of the
MBBS course is five years and six months, including one year of rota-
tional internship
5 ANM or Auxiliary Nursing Midwifery is a two-year diploma programme
in the field of Nursing
6 Aarogya Setu is the official mobile application for COVID-19 developed
by the National Informatics Centre under the Ministry of Electronics and
Information Technology, Government of India

REFERENCES
Acharjee, S. (2021). India’s covid collapse, part 1: How Modi government’s
complacency in keeping track of new mutants triggered a second wave.
Retrieved May 30, 2021, from https://www.indiatoday.in/magazine/
cover-story/story/20210517-india-s-covid-collapse-part-1-how-
modi-government-s-complacency-in-keeping-track-of-new-mutants-
triggered-second-wave-1800992-2021-05-10
Bajekal, N. (2021). How did India’s Covid-19 Crisis become a catastrophe?
Retrieved May 30, 2021, from https://time.com/5964796/india-
covid-19-failure/
Barney, B. J., & Schulzke, K. S. (2016). Moderating “cry wolf” events with
excess MAD in Benford’s law research and practice. Journal of Forensic
Accounting Research, 1(1), A66–A90.
Benford, F. (1938). The law of anomalous numbers. Proceedings of the
American Philosophical Society, 78, 551–572.
Bhattacharya, S. (2020). Covid-19 deaths in Bengal double in 5 days as govt
junks casualty audit panel. Retrieved May 30, 2021, from https://www.
hindustantimes.com/india-news/covid-19-deaths-in-bengal-double-
in-5-days-as-govt-junks-casualty-audit-panel/story-RFvtl7g3WPH5GR
bdOQbQM.html
Carslaw, C. A. (1988). Anomalies in income numbers: Evidence of goal
oriented behavior. Accounting Review, 63(2), 321–327.
Dasgupta, S. (2021). Covid testing drops in Delhi, labs say they don’t have
enough staff, skilled technicians. The Print. Retrieved May 28, 2021,
from https://theprint.in/health/covid-testing-drops-in-delhi-labs-say-
they-dont-have-enough-staff-skilled-technicians/646902/
Desai, D. (2020). Gujarat govt is managing data, not coronavirus, say medi-
cal Experts: Percentage of deaths in state among highest, even as num-
ber of tests remain low. Firstpost. Retrieved May 28, 2021, from
https://www.firstpost.com/india/gujarat-govt-is-managing-data-not-
coronavirus-say-medical-experts-percentage-of-covid-19-deaths-in-
state-among-highest-even-as-number-of-tests-remain-low-8479721.
html
Drake, P. D., & Nigri, M. J. (2000). Computer assisted analytical procedures
using Benford’s law. Journal of Accounting Education, 18(2), 127–146.
Druica, E., Oancea, B., & Vălsan, C. (2018). Benford’s law and the limits of
digit analysis. International Journal of Accounting Information Systems,
31, 75–82.
Ethiraj, G. (2021). By hiding the real number of COVID-19 cases and Deaths,
some Indian states are Disempowering people. Retrieved May 29, 2021,
from https://scroll.in/article/993324/by-hiding-the-real-number-of-covid-
19-cases-and-deaths-some-indian-states-are-disempowering-people
Gettleman, J., Yasir, S., Kumar, H., & Raj, S. (2021). As Covid-19 devastates
India. In Deaths go uncounted. New York (USA): The New York
Times Company. Retrieved May 29, 2021, from https://www.nytimes.
com/2021/04/24/world/asia/india-coronavirus-deaths.html
Gonzalez-Garcia, J., & Pastor, G. (2009). Benford’s Law and Macroeconomic
Data Quality. Washington (USA): International Monetary Fund Work-
ning Paper, 2009.
Gonzalez, F. A. I. (2020). Self-reported income data: Are people telling the
truth? Journal of Financial Crime., 27(4), 1349–1359.
Henselmann, K, Scherr, E, & Ditter, D. (2013). Applying Benford’s Law to
individual financial reports: An empirical investigation on the basis of SEC
XBRL filings [Working Paper No. 2012-1] (rev.1.) Working Papers in
Accounting Valuation Auditing. https://www.econstor.eu/handle/
10419/88418
Johnson, G. G. (2005). Financial sleuthing using Benford’s law to analyze
quarterly data with various industry profiles. Journal of Forensic
Accounting, 6(2), 293–316.
Judge, G., & Schechter, L. (2009). Detecting problems in survey data using
Benford’s law. Journal of Human Resources, 44(1), 1–24.
Koch, C., & Okamura, K. (2020). Benford’s law and COVID-19 reporting.
Economics Letters, 196, 109573. https://doi.org/10.1016/j.econlet.
2020.109573
Mallapaty, S. (2021). India’s massive COVID SURGE Puzzles scientists. Retrieved May 28, 2021, from https://www.nature.com/articles/d41586-021-01059-y

Mir, T. A., Ausloos, M., & Cerqueti, R. (2014). Benford’s law predicted digit distribution of aggregated income taxes: The surprising conformity of Italian cities and regions. The European Physical Journal B, 87(11), 1–8.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. American Journal of Mathematics, 4(1), 39–40.

Nigrini, M. J. (1996). A taxpayer compliance application of Benford’s law. The Journal of the American Taxation Association, 18(1), 72.

Nigrini, M. J. (2005). An assessment of the change in the incidence of earnings management around the Enron-Andersen episode. Review of Accounting and Finance, 4(1), 92–110.

Nye, J., & Moul, C. (2007). The political economy of numbers: On the application of Benford’s law to international macroeconomic statistics. The BE Journal of Macroeconomics, 7(1), 1–14.

Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50(302), 157–175. https://doi.org/10.1080/14786440009463897

Pulla, P. (2021). There are so many hurdles. Indian scientists plead with government to unlock COVID-19 data. Retrieved May 30, 2021, from https://www.sciencemag.org/news/2021/05/there-are-so-many-hurdles-indian-scientists-plead-government-unlock-covid-19-data

Raj Kumar, S, & Devakumar, K. (2021). COVID-19 in India: Dataset on Novel Corona Virus Disease 2019 in India. https://kaggle.com/sudalairajkumar/covid19-in-india

Ramakrishnan, S. (2021). The meta bloc: Why India faces a data shortage on genome sequencing of coronavirus. Retrieved May 30, 2021, from https://www.newindianexpress.com/nation/2021/may/02/the-meta-blocwhy-india-faces-a-data-shortage-on-genome-sequencing-of-coronavirus-2297425.html

Ramani, S, & Radhakrishnan, V (2021, May 29). Kolkata’s COVID-19 deaths in 2021 could be 4 times higher. Retrieved May 30, 2021, from https://www.thehindu.com/news/cities/kolkata/kolkatas-covid-19-deaths-in-2021-could-be-4-times-higher/article34671165.ece

Rauch, B., Göttche, M., Engel, S., & Brähler, G. (2011). Fact and fiction in EU-governmental economic data. German Economic Review, 12(3), 243–255.

Sharma, M. (2021, April 07). Data shows drastic fall in COVID-19 tests is helping Bengal. In Assam keep their numbers low. New Delhi (India): Living Media India Limited. Retrieved May 30, 2021, from https://www.indiatoday.in/coronavirus-outbreak/story/data-shows-drastic-fall-in-covid-19-tests-is-helping-bengal-assam-keep-their-numbers-low-1788219-2021-04-07

Thomas, J. K. (1989). Unusual patterns in reported earnings. Accounting Review, 64(4), 773.

Tumbe, C. (2021). Why ‘excess mortality’ figures for covid must be calculated. Retrieved May 29, 2021, from https://indianexpress.com/article/opinion/columns/why-excess-mortality-figures-for-covid-must-be-calculated-7330348/

Villas-Boas, S. B., Fu, Q., & Judge, G. (2017). Benford’s law and the FSD distribution of economic behavioral micro data. Physica A: Statistical Mechanics and its Applications, 486, 711–719.

How to cite this article: Mahasuar, K. (2021). Lies, damned lies, and statistics: The uncertainty over COVID-19 numbers in India. Knowledge and Process Management, 1–8. https://doi.org/10.1002/kpm.1685