JJMS **18**, No. 2, 261-274 (2025)

Jordan Journal of Mathematics and Statistics.
Yarmouk University

DOI:https://doi.org/10.47013/18.2.10

How Dependable is COVID-19 Data During First Wave? Disclosure of Inconsistencies in Daily Reportage Confirmed Cases and Deaths

Edmund F. Agyemang[1,2,3*], *Joseph A. Mensah*[3], *Eric Nyarko*[1], *Ezekiel N.N. Nortey*[1]

[1] Department of Statistics and Actuarial Science, College of Basic and Applied Sciences, University of Ghana
[2] School of Mathmatics and Statistical Science, College of Sciences, University of Texas Rio Grande Valley, USA
[3] Department of Computer Science and Information Systems, Ashesi University, No.1 University Avenue, Berekuso-Ghana

Received: Feb. 27, 2024 Accepted : Mar. 19, 2025

Abstract: The global crisis triggered by the COVID-19 pandemic necessitated precise data monitoring and rigorous analysis efforts from worldwide health authorities and governments, particularly during the pandemic's initial surge. This study employs Newcomb Benford's Law specifically to identify potential anomalies in the reporting of COVID-19 data during the pandemic's first wave. Our methodology encompasses the application of Newcomb Benford's Law to the first digit analysis focusing on three (3) key statistics [d^*, $\alpha-$ statistic (α^*) and ω statistic (ω^*)] in conjunction with the Kolmogorov-Smirnov test to unveil possible inconsistencies within world continental COVID-19 data reportage. By evaluating the actual distribution of leading digits across various COVID-19 data categories such as cumulative confirmed cases, deaths, recoveries, and active cases against the theoretical distribution proposed by Newcomb Benford's Law, possible significant deviations were identified. We used the deviation from the Newcomb Benford's law of anomalous numbers as a proxy for data accuracy. The findings reveal that except for the Australia/Oceania continent which exhibited pronounced deviations due to its unique data structure, the COVID-19 data from all other continents maintained a possible high level of reliability during the initial outbreak. The study concludes that while Benford's Law is a valuable tool for anomaly detection in diverse data, its use in COVID-19 reportage data shows potential pitfalls. To enhance the effectiveness and reliability of detecting anomalies, the study advocates for integrating additional anomaly detection strategies, like density and boundary based approaches encompassing the local outlier factor and one-class SVM, alongside the Newcomb Benford analysis.

Keywords: COVID-19; Newcomb Benford's Law; First-digits; World Continents; Kolmogorov-Smirnov test.

1 Introduction

The World Health Organization (WHO) declared the COVID-19 outbreak a Public Health Emergency of International Concern (PHEIC) on January 30, 2020 [1]. The WHO issued this declaration in response to COVID-19 spreading rapidly outside of China and its potential threat to other nations. To contain the spread of COVID-19, the WHO declared COVID-19 a PHEIC to mobilize countries and communities to act urgently and aggressively. A PHEIC declaration allows WHO to provide technical and operational support to countries [2]. Additionally, it provides a means of coordinating international efforts to battle outbreaks and mobilize resources. Furthermore, it advocated joint coordination and information sharing among participating countries so that the spread of the virus could be minimized and effective responses could be implemented [3]. On May 5, 2023, the WHO ended its declaration of COVID-19 being a global heath emergency but continued to refer to it as a pandemic. Since COVID-19 became a pandemic, almost every country adopted a comprehensive testing program and reported positive cases. Data on new cases and deaths reported by health authorities on a daily basis was vital for analyzing disease dynamics.

* Corresponding author e-mail: edmundfosu6@gmail.com, ORCID ID: 0000-0001-8124-4493

The impact of COVID-19 was felt globally, with far-reaching consequences across economies and disciplines. Many of the world's economies experienced shocks: businesses closing down, increasing unemployment rates, reduction in global trade, disruptions in supply chains, and several loses in the service and manufacturing sector [4,5]. The COVID-19 outbreak significantly impacted worldwide educational systems, including West African countries, driving schools to close campuses [5]. School closures due to the pandemic affected more than 1 billion students worldwide. The pandemic also led to a shift in the deployment of online learning, which resulted in uneven access to educational resources, exacerbating existing educational inequalities [6]. Moreover, the closure of schools has had a significant impact on the well-being of children and their families [7]. Furthermore, the COVID-19 pandemic had a significant impact on the tourism industry, with travel restrictions and border closures leading to decreased demand and revenue loss [8,9]. The pandemic has also led to a shift towards domestic tourism and an increased focus on sustainable tourism [10, 11]. The COVID-19 pandemic likewise highlighted existing health disparities and inequities, with marginalized populations and low-income countries being disproportionately affected [12,13]. These disparities are linked to social determinants of health, including poverty, discrimination, and inadequate access to healthcare [14,15]. The pandemic has shed light on the importance of addressing these underlying factors to ensure equitable health outcomes for all.

To mitigate the impact of the pandemic, governments worldwide successfully employed various statistical approaches to inform their decision-making. These spanned techniques aimed to forecast the spread of COVID-19 and its impact on health care systems [16,17,18,19], determine the effectiveness of different mitigating strategies, such as social distancing and mask mandates [20,21], monitoring the effectiveness of COVID-19 vaccines through clinical trials and real-world data analysis [22,23,24,25].

Nonetheless, it is essential to note that several political and economic factors could influence the registration and reporting of COVID-19 cases and deaths. Due to this, the reported data may be prone to errors, resulting in a biased analysis of the spread of the disease [26]. This is due to the fact that some health authorities are likely to be influenced by political pressures to under-report COVID-19 cases in order to calm the nerves of their citizens but this also has its own implications in the long run. Besides, several researchers reported of many countries' COVID-19 data on daily cases and deaths having varying levels of accuracy and completeness due to challenges such as limited testing capacity, and overwhelmed healthcare systems [27]. For example, many studies analyzing COVID-19 data from several continents found evidence of under-reporting of COVID-19 cases and deaths, likely due to factors such as insufficient testing and political interference in reporting: South America [28,29], Asia [30], Africa [31,32,33] and worldwide [34,35,36,37]. By not accurately communicating the extent of COVID-19 cases, individuals may underestimate the risk and fail to take appropriate precautions, and this can result in more severe health outcomes and burden the healthcare system by increasing the chances of virus transmission.

To scrutinize reported data, researchers have advocated for the use of anomaly detection techniques. Anomaly detection techniques such as Newcomb Benford's Law and Bayesian analysis have been employed practically and technically to detect possible fraud or anomalies in most real-world datasets. Their applicability encompasses detecting possible or suspected fraud in welfare programs [38], detection of fraud in accounting [39], image forensics [40], election anomalies [41,42,43,44] and others. Newcomb-Benfors's Law examines the distribution of first digits within datasets, enabling statisticians to detect any discrepancies that might allude to deceitful or falsified figures [45]. For example, if the first digit distribution of COVID-19 case counts deviates significantly from what is postulated by Newcomb Benford's Law, it could suggest that the data has been tampered with. While using Benford's Law in fraud detection is not foolproof and should be used in conjunction with other methods, it can be an effective tool for identifying potential irregularities in COVID-19 data and warrants further investigation [26].

The study employs Newcomb Benford's Law to the initial digit analysis, focusing on three (3) key statistics [d^*, $\alpha-$ statistic (α^*), and ω statistic (ω^*)], in addition to the Kolmogorov-Smirnov test (defined in Section 2.1), to reveal inconsistencies within continental COVID-19 data. By evaluating the actual distribution of leading digits across various COVID-19 data categories—such as cumulative confirmed cases, deaths, recoveries, and active cases—against the theoretical distribution proposed by Newcomb Benford's Law, we were able to identify significant deviations. COVID-19 data collection, compilation and reporting may be assisted by Benford's law as an efficient screening method, hence its applicability in the present study. The remainder of the paper is organized as follows: Section 2 discusses the data and methods used for the study which include goodness of fit measures for Newcomb Benford's Law. Section 3 presents and discusses the results of the investigation. Section 4 concludes the research and provides recommendations for further work. Section 5 contains the appendix.

2 Data and Methods

The study made use of secondary data of cumulative confirmed cases, cumulative confirmed deaths, cumulative confirmed recovered cases and cumulative active cases between December 30, 2019 and August 4, 2020 (with data collected during the first wave of the COVID-19 pandemic) retrieved from the website of John Hopkins University in the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) and the website of World Health Organization. These data sources can be assessed via https://coronavirus.jhu.edu/region and https://data.who.int/dashboards/covid19/data. Initially, the COVID-19 dataset was filtered out to remove all countries that had less than 1,000 cases for most of the characteristics that were under consideration. The exception was for countries in the Australia/Oceania region. This requirement was selected because data sets that span more than three orders of magnitude are more likely to conform to Benford's law than data sets spanning only one or two orders of magnitude [46]. A function for the extraction of the first digits observed and expected proportions of cumulative confirmed cases, cumulative confirmed deaths, cumulative confirmed recovered cases and cumulative active cases was implemented with the help of R programming statistical software. Next, Benford's Law was applied to the first digits of the selected metrics, and the expected distribution was calculated continental-wise. Thus, the expected distribution for a specific continent is computed by combining all countries that form part of the continent. Finally, euclidean distance variations [47], such as the [48] d^*, [49] Statistic (α^*), [50] Statistic (ω^*) and statistical hypothesis test such as Kolmogorov-Smirnov test were employed to evaluate the degree of conformity between the observed and Benford's expected distributions of COVID-19 reportage cases and deaths of all world continental data.

2.1 Goodness-of-fit measures for Benford's Law

To measure how well the data comply with the distribution posited by Benford Law, the following goodness-of-fit measures were considered.

(a) d^*: This quantifies the deviation from Benford's Law [48,51]. The d^* is the euclidean distance between each world continental probable COVID-19 cases and Benford's posited/expected COVID-19 cases after standardization by the maximum possible distance of 1.03606, the situation when there is only a peak at "9" and zero for other first digits.

The d^* is given in (1) by

$$d^* = \frac{\sqrt{\sum_{d_i=1}^{9}\left[P(g_i) - \tilde{P}(g_i)\right]^2}}{1.03606}, \tag{1}$$

where $d_i \in (1,2,\ldots,9)$, $P(g_i)$ are the observed proportions of the first digits and $\tilde{P}(g_i)$ are the Benford's postulated proportions of the digits. A d^* value of 0.0 indicates the exactness matching of a world continental COVID-19 data (cumulative confirmed cases, cumulative confirmed deaths, cumulative confirmed recovered and cumulative confirmed active cases) to the Benford's curve; possible non-anomalous COVID-19 data reported. A higher d^* value implies lower "Benfordness," thus d^* closer to 1.0. A d^* value higher than 0.25 implies higher "Non-Benfordness," thus high evidence of possible data manipulation [51]. Benford's Law is justified for detecting possible anomalies of COVID-19 data with a low world d^* of 0.03.

(b) Chi-Square Statistic, χ^2_{stat}

$$\chi^2_{stat} = n \cdot \sum_{j=10^{k-1}}^{10^k-1} \frac{\left(f_j^o - f_j^e\right)}{f_j^e}, \tag{2}$$

where f_j^o and f_j^e are respectively the observed and expected Benford frequencies of digits j. However, for small sample size, the χ^2_{stat} in (2) has low statistical power relative to its enormous statistical power when used for larger sample size [62].

(c) An alternative good of fit measure, less dependent on sample size n, a modified version of the [65] proposed by [64] and [63] is recommended for use in extant literature defined in (3).

$$Kupier = (D^+ + D^-)\left[\sqrt{n} + 0.155 + \frac{0.24}{\sqrt{n}}\right], \tag{3}$$

$$D^+ = \sup_{-\infty<x<+\infty}\left[F_o(x) - F_e(x)\right],$$

and

$$D^{-} = \sup_{-\infty<x<+\infty} [F_e(x) - F_o(x)],$$

where $F_o(x)$ is the observed cumulative density function (CDF) of leading digits and $F_e(x)$ is the CDF of Benford's data.

(d) Kolmogorov-Smirnov Test

$$K_{\phi}^{*} = \sup_{i=10^{k-1},\ldots,10^{k}-1} \left\| \sum_{j=1}^{i} \left(f_j^o - f_j^e\right) \right\| \cdot \sqrt{n}, \tag{4}$$

where f_j^o and f_j^e are defined as above. Here, the null hypothesis is rejected if the test statistic, given by (4) exceeds the critical values d, as provided by [61] and [60] given in (5) as;

$$d = \frac{\gamma_\kappa}{\sqrt{n}}, \tag{5}$$

where $\gamma_\kappa \in \{1.224, 1.358, 1.628\}$ depends on the level of significance. Central to this study were statistical decisions on the Kolmogorov-Smirnov test based on the p-value. Thus, a p-value greater than an indicated α level of 0.05 represents possible "Benfordness" whilst a p-value less than an indicated α level of 0.05 represents possible "Non-Benfordness". With 10,000 iterations, the "simulate" method was used in calculating the p-value of the Kolmogorov-Smirnov test.

(e) Furthermore, both [49] and [50] proposed a distance measure criterion [the $\omega-$ statistic (ω^*) and the $\alpha-$ statistic (α^*) respectively] to test the degree of similarity between the distribution of the digits analyzed and that of the Newcomb Benford expected distribution expressed in (6) and (7) as

$$\omega^* = \max_{g=1}^{9} \left\| f_j^o - f_j^e \right\|, \tag{6}$$

and

$$\alpha^* = \sqrt{\sum_{g=1}^{9} (p_j^o - p_j^e)^2}, \tag{7}$$

where p_j^o are the observed proportions of observations with first digits g and p_j^e are the corresponding Benford postulated proportions of first digits g. The ω^* and α^* statistics are less sensitive to sample size. Similar to the decision criterion of the d^*, possible "Benfordness" and "Non-Benfordness" of both ω^* and α^* Euclidean distance measures were decided accordingly.

For the purpose of this study, we report the results of the d^*, $\omega-$ statistic (ω^*), $\alpha-$ statistic (α^*) and the Kolmogorov-Smirnov test statistic (K_ϕ^*).

The marginal probability mass function (PMF) of the first digits (G_1) is given in (8) by:

$$Pr(G_1 = g_1) = \log\left(\frac{g_1+1}{g_1}\right), \quad g_1 \in (1,2,\ldots,9). \tag{8}$$

The probability distribution for each number from 1 to 9 to be the first digit according to Benford's Law is shown in Table 1

Table 1: First digits distribution according to Benford's Law

Digits, g_1	1	2	3	4	5	6	7	8	9
$P(g_1)$	0.3010	0.1761	0.1249	0.0969	0.0792	0.0669	0.0580	0.0512	0.0452

This law was expanded by [59] to include digits other than the first given in (9) as follows;

$$P(G_1 = g_1, \ldots, G_k = g_k) = \log_{10}\left(1 + \frac{1}{\sum_{i=1}^{k} 10^{k-1} g_i}\right), \tag{9}$$

where $g_1 \in \{1, \ldots, 9\}$ and all other $g_j \in \{0, \ldots, 9\}$ for $j = 2, \ldots, k$.

For the purpose of this study, in the case where there is a conflicting conclusion (either possible Benfordness or non-Benfordness) between two or more goodness of fit tests, the general criterion to conclude is defined in (10) as follows. Let P_ϕ^*: P-values of the Kolmogorov-Smirnov test. We define a criterion to decide between possible Benfordness and non-Benfordness based on these values:

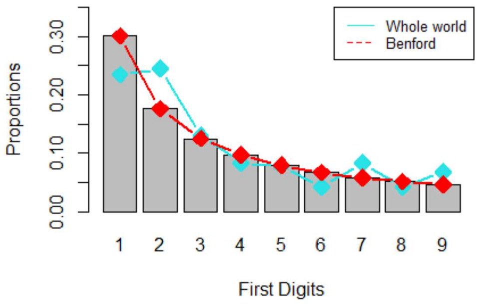
$$Criterion = \begin{cases} \text{If at least two of } d^*, \alpha^* \text{ and } \omega^* \leq 0.25 \text{ and } P_\phi^* \geq 0.05; & \text{Possible Benfordness;} \\ Otherwise; & \text{Possible Non-Benfordness.} \end{cases} \tag{10}$$

3 Results and Discussion

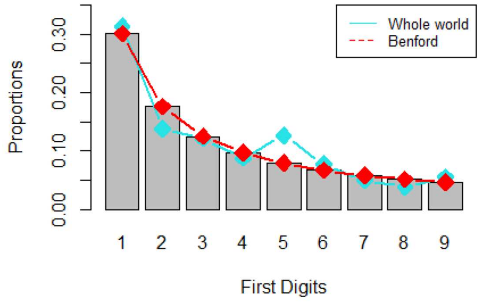
The empirical findings from the study are presented in this section.

3.1 Justification for the use of Benford's Law

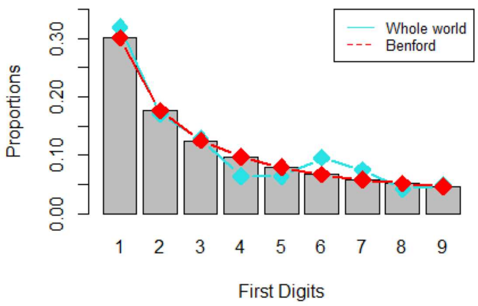
The application of Newcomb Benford's Law to assess world continental COVID-19 data was explored in this study. The d^*, α^* and ω^* values of the whole world data were approximately below 0.10. The results of the confirmatory analysis are shown in the last row of Table 2, Table 3, Table 4 and Table 5. Moreover, the histogram showing the frequency distribution depicts how well world continental COVID-19 data fits into the Newcomb Benford's curve with little or no deviation from the Benford's posited distribution [see Figures 1(a), 1(b), 1(c) and 1(d)]. The p-values of the Kolmogorov-Smirnov test are all greater than 0.05 indicating additional support for the use of Newcomb Benford's Law analysis for the anomaly detection of world continental COVID-19 data. These results validate the use of Newcomb Benford's Law for the analysis of world continental COVID-19 data, in all the four characteristics of interest (cumulative confirmed cases, cumulative deaths, cumulative recorded cases and cumulative active cases) under consideration. The correlation between the three Euclidean measures (d^*, α^* and ω^*) were computed by fitting the first digit distribution of the four variables (cumulative confirmed cases, cumulative confirmed deaths, cumulative recovered cases and cumulative active cases) of studied world continents into the Benford's distribution. The blue line in Figure 6(a), Figure 6(b), and Figure 6(c) in the appendix traces the linear regression. With coefficient of determination (R^2) values of 1.0, 0.9547 and 0.9547 as shown at the top right corner in Figure 6(a), Figure 6(b), and Figure 6(c), the results suggest that d^*, α^* and ω^* combine to provide best fit to the cumulative world continental COVID-19 data.



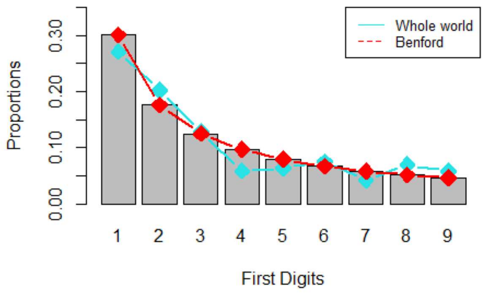
(a) Distribution of first digits cumulative confirmed cases: Whole World



(b) Distribution of first digits cumulative confirmed deaths: Whole World



(c) Distribution of first digits cumulative recovered cases: Whole World



(d) Distribution of first digits cumulative active cases: Whole World

Fig. 1: First digits distribution between Benford and Whole World cumulative continental COVID-19 cases

The objective of this study is to gather meaningful insights through a comprehensive scrutiny of the outcomes obtained using the Newcomb Benford's law to evaluate the accuracy and reliability of COVID-19 data. By using Benford's Law to analyze COVID-19 data from different continents, this study gave room for comparative analysis to be conducted on data reported by world continents during the first wave of the COVID-19 pandemic. Three prominent scenarios emerged: adherence to Benford's law principles; deviations thereof; or ambiguity surrounding such findings altogether. Instances, where conformity existed per these principles, increased confidence levels regarding legitimacy and reliability within COVID-19-related datasets analyzed are highlighted. Even though a good accordance and promising degree of agreement between [48] d^*, [49] Statistic (α^*), [50] Statistic (ω^*) and Kolmogorov-Smirnov statistic (K_ϕ^*) of the analysis of data conformity of world continental COVID-19 related variables (cumulative confirmed cases, cumulative confirmed deaths, cumulative recovered cases and cumulative active cases) to the Benford's Law could be observed, some divergence were noted for these four (4) analytical tools presented in this study. For example, data from the Australia/Oceania continent presented d^*, α^* and ω^* values above the threshold of 0.25 for all the variables under consideration though the K_ϕ^* test failed to reject the null hypothesis of possible conformity to Benford's Law (with $P_\phi^* > 0.05$) as evident in Table 2, Table 3, Table 4 and Table 5.

On the other hand, data from the South American continent (cumulative confirmed COVID-19 cases) reported d^* and α^* values above the Benford conformity threshold of 0.25 with a ω^* value of 0.2177 less than 0.25 signalling possible conformity to Newcomb Benford's Law and a P_ϕ^* value of 0.2661 greater than 0.05 produced by the Kolmogorov-Smirnov test indicating possible conformity to Benford's Law. Using the criterion established in (10), the cumulative COVID-19 confirmed cases for South America continent and all COVID-19 related variables of

Deanship of Research and Graduate Studies, Yarmouk University, Irbid, Jordan.

Austrailia/Oceania continent were adjudged not to conform to the distributional pattern postulated by Newcomb Benford's Law. Thus, data from these sources may be prone to possible data reportage issues including but not limited to data manipulation, data input issues, and data reporting errors and are therefore subject to further scrutiny.

The results of this study suggest that d^*, α^*, ω^* and K^*_ϕ converge to act together as preliminary statistical tools to evaluate the conformity of world continental COVID-19 data to the Newcomb Benford's Law during the first wave of the COVID-19 pandemic. Furthermore, this work assessed the performance of two other distance measures namely [49] Statistic (α^*) and [50] Statistic (ω^*) in addition to the originally used [48] statistic (d^*) to check for reliability of COVID-19 data. The argument that the Euclidean distance is not a hypothesis test as argued by other researchers was addressed accordingly in this work. In this study, the authors resort to the Kolmogorov-Smirnov Test for Benford's Law which is a hypothesis test to confirm the degree of COVID-19 data accuracy reported by the various continents of the world during the first wave of the COVID-19 pandemic. A striking obseervation of the three euclidean measures considered was that the [48] statistic (d^*) and the [49] Statistic (α^*) were found to produce approximately same Euclidean distance values.

Table 2: d^*, ω^* Statistic, α^* Statistic and Kolmogorov-Smirnov K^*_ϕ and its corresponding P-value (P^*_ϕ) results calculated from the first digit distribution of World Continental cumulative confirmed COVID-19 cases fitted into the Benford's distribution

	Test Statistic					Statistical Decision			
Continent	d^*	$\alpha-$ Stat. (α^*)	$\omega-$ Stat. (ω^*)	(K^*_ϕ)	P^*_ϕ	d^*	α^*	ω^*	K^*_ϕ
Africa	0.0738	0.0765	0.0421	0.2482	0.9678	✓	✓	✓	✓
Asia	0.1217	0.1260	0.0792	0.4979	0.6810	✓	✓	✓	✓
Australia/Oceania	0.4279	0.4434	0.2990	0.8725	0.1940	×	×	×	✓
Europe	0.1516	0.1570	0.1106	0.7710	0.2902	✓	✓	✓	✓
North America	0.1738	0.1801	0.0959	0.5423	0.6650	✓	✓	✓	✓
South America	0.2935	0.3041	0.2177	0.8080	0.2661	×	×	✓	✓
Whole World	0.1023	0.1060	0.0687	0.9429	0.1323	✓	✓	✓	✓

d^*: [48]; α^*: [49] Statistic; ω^*: [50] Statistic; K^*_ϕ: Kolmogorov-Smirnov Statistic; P^*_ϕ: P-values of the Kolmogorov-Smirnov test; ✓ represent possible Benfordness and × represent possible Non-Benfordness.

Table 3: d^*, ω^* Statistic, α^* Statistic and Kolmogorov-Smirnov K^*_ϕ and its corresponding P-value (P^*_ϕ) results calculated from the first digit distribution of World Continental cumulative COVID-19 deaths fitted into the Benford's distribution.

	Test Statistic					Statistical Decision			
Continent	d^*	$\alpha-$ Stat. (α^*)	$\omega-$ Stat. (ω^*)	(K^*_ϕ)	P^*_ϕ	d^*	α^*	ω^*	K^*_ϕ
Africa	0.1184	0.1227	0.1095	0.7817	0.2706	✓	✓	✓	✓
Asia	0.1716	0.1778	0.1273	0.5071	0.6955	✓	✓	✓	✓
Australia/Oceania	0.5403	0.5598	0.4906	0.7959	0.2501	×	×	×	✓
Europe	0.1004	0.1040	0.0760	0.5362	0.6094	✓	✓	✓	✓
North America	0.2337	0.2421	0.1990	0.9377	0.1393	✓	✓	✓	✓
South America	0.2180	0.2258	0.1343	0.5307	0.6826	✓	✓	✓	✓
Whole World	0.0640	0.0663	0.0472	0.5848	0.5478	✓	✓	✓	✓

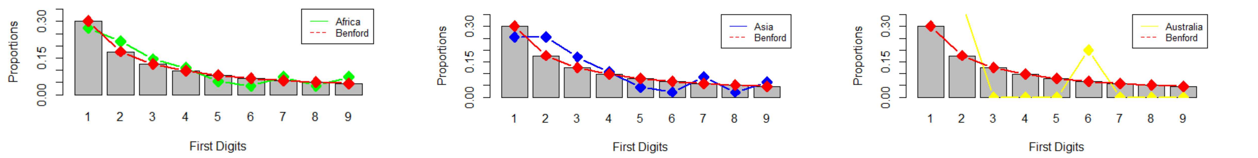
Table 4: d^*, ω^* Statistic, α^* Statistic and Kolmogorov-Smirnov K^*_ϕ and its corresponding P-value (P^*_ϕ) results calculated from the first digit distribution of World Continental cumulative recovered COVID-19 cases fitted into the Benford's distribution.

	Test Statistic					Statistical Decision			
Continent	d^*	$\alpha-$ Stat. (α^*)	$\omega-$ Stat. (ω^*)	(K^*_ϕ)	P^*_ϕ	d^*	α^*	ω^*	K^*_ϕ
Africa	0.1322	0.2555	0.1370	0.0808	0.0589	✓	✓	✓	✓
Asia	0.1350	0.1399	0.1005	0.4979	0.6946	✓	✓	✓	✓
Australia/Oceania	0.3804	0.3941	0.2990	0.4874	0.6470	×	×	×	✓
Europe	0.1392	0.1442	0.0708	0.3760	0.8766	✓	✓	✓	✓
North America	0.1714	0.1775	0.0954	0.4341	0.7308	✓	✓	✓	✓
South America	0.1978	0.2049	0.1156	0.3254	0.9058	✓	✓	✓	✓
Whole World	0.0518	0.0537	0.0331	0.5143	0.6637	✓	✓	✓	✓

Table 5: d^*, ω^* Statistic, α^* Statistic and Kolmogorov-Smirnov K^*_ϕ and its corresponding P-value (P^*_ϕ) results calculated from the first digit distribution of World Continental cumulative active COVID-19 cases fitted into the Benford's distribution.

	Test Statistic					Statistical Decision			
Continent	d^*	$\alpha-$ Stat. (α^*)	$\omega-$ Stat. (ω^*)	(K^*_ϕ)	P^*_ϕ	d^*	α^*	ω^*	K^*_ϕ
Africa	0.1191	0.1234	0.0966	0.3819	0.8484	✓	✓	✓	✓
Asia	0.1586	0.1643	0.1091	0.5632	0.5483	✓	✓	✓	✓
Australia/Oceania	0.4560	0.4725	0.3488	0.9574	0.0807	×	×	×	✓
Europe	0.0968	0.1003	0.0524	0.5149	0.6934	✓	✓	✓	✓
North America	0.1887	0.1955	0.1027	0.8436	0.1904	✓	✓	✓	✓
South America	0.2417	0.2500	0.1990	0.5787	0.5144	✓	✓	✓	✓
Whole World	0.0615	0.0637	0.0384	0.9226	0.1630	✓	✓	✓	✓

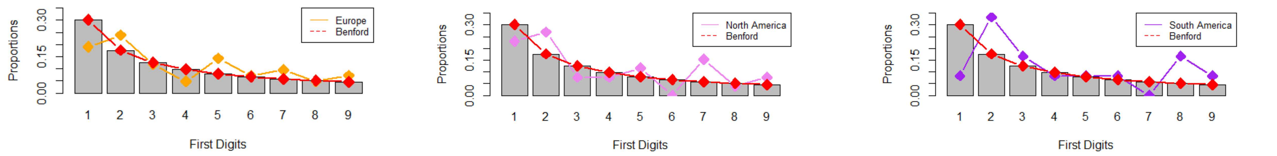
Here, we apply Benford's first digits distributional pattern to the world continental COVID-19 data for cumulative confirmed cases, cumulative deaths, cumulative recovered cases and cumulative active cases as shown in Figures 2, 3, 4 and 5. Based on the distribution of first digits, Figure 2(c), 3(c), 4(c) and 5(c) indicate that most observed proportions do not conform to Benford's posited proportions and hence Australia/Oceania continent display severe possible nonconformity to Newcomb Benford's law which is subject to further investigation. The distributional pattern of first digits for South America continent's cumulative confirmed cases [Figure 2(f)], cumulative recovered cases [Figure 4(f)], cumulative confirmed deaths [Figure 3(f)] and North America's cumulative recovered cases [Figure 4(e)] exhibited minimal divergence from the Benford's fitted curve. Nearly perfect conformity to Newcomb Benford's Law are observed for Africa's cumulative confirmed cases [Figure 2(a)], cumulative confirmed deaths [Figure 3(a)] and cumulative active cases [Figure 5(a)] and Asia's cumulative confirmed cases [Figure 2(b)], cumulative recovered cases [Figure 4(b)]. Likewise, near conformity to Benford's Law holds for European continent's cumulative confirmed cases, cumulative confirmed deaths, cumulative recovered cases and cumulative active cases as evident in Figures 2(d), 3(d),4(d), and 5(d) respectively.



(a) Distribution of first digits cumulative confirmed cases: Africa

(b) Distribution of first digits cumulative confirmed cases: Asia

(c) Distribution of first digits cumulative confirmed cases: Australia

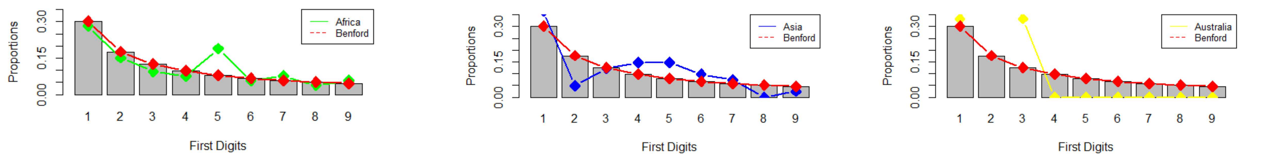


(d) Distribution of first digits cumulative confirmed cases: Europe

(e) Distribution of first digits cumulative confirmed cases: North America

(f) Distribution of first digits cumulative confirmed cases: South America

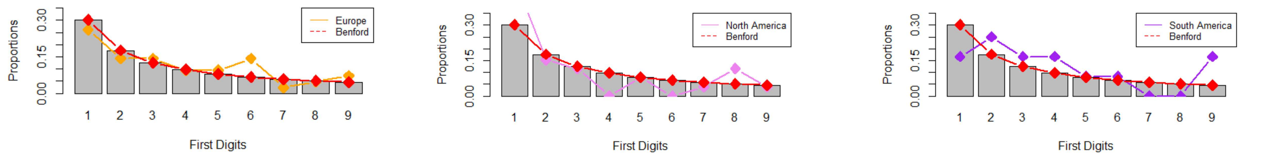
Fig. 2: First digits distribution between Benford and World continental cumulative confirmed COVID-19 cases



(a) Distribution of first digits cumulative confirmed deaths: Africa

(b) Distribution of first digits cumulative confirmed deaths: Asia

(c) Distribution of first digits cumulative confirmed deaths: Australia



(d) Distribution of first digits cumulative confirmed deaths: Europe

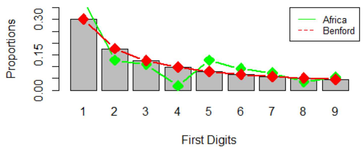
(e) Distribution of first digits cumulative confirmed deaths: North America

(f) Distribution of first digits cumulative confirmed deaths: South America

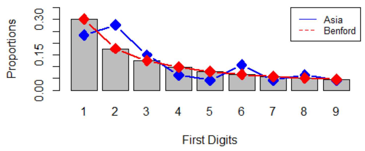
Fig. 3: First digits distribution between Benford and World continental cumulative confirmed COVID-19 deaths

3.2 The case of Australia/Oceania continent

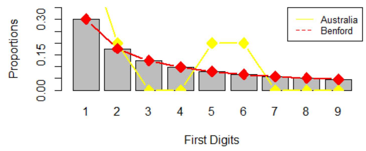
Benford's Law when however, applied to small samples, the distribution of leading digits may deviate significantly from the expected distributional pattern postulated by Benford [58]. As a result, there can be deviations from the expected distribution in small samples due to the limited number of data points and the inherent randomness [57]. It is therefore important to take into consideration the potential bias of limited data when applying Benford's Law to small sets of data, even though it can be useful when analyzing large datasets as noted by [26]. Based on the aforementioned, it is prudent to argue that the C0VID-19 data reportage of Australia/Oceania-related variables is not completely fraudulent



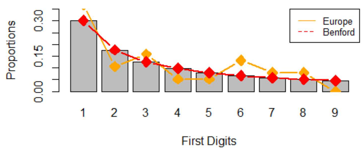
(a) Distribution of first digits cumulative recovered cases: Africa



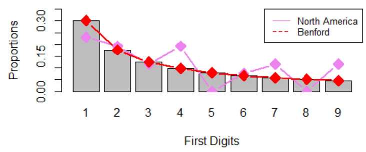
(b) Distribution of first digits cumulative recovered cases: Asia



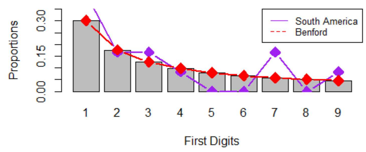
(c) Distribution of first digits cumulative recovered cases: Australia



(d) Distribution of first digits cumulative recovered cases: Europe

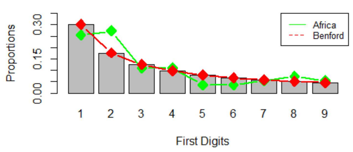


(e) Distribution of first digits cumulative recovered cases: North America

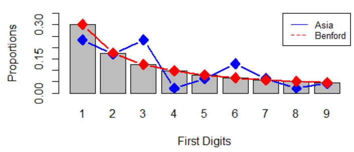


(f) Distribution of first digits cumulative recovered cases: South America

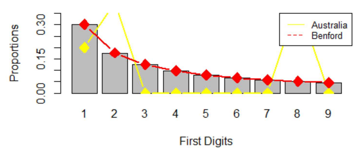
Fig. 4: First digits distribution between Benford and World continental cumulative confirmed recovered COVID-19 cases



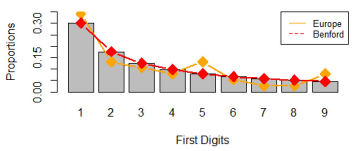
(a) Distribution of first digits cumulative active cases: Africa



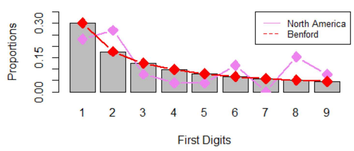
(b) Distribution of first digits cumulative active cases: Asia



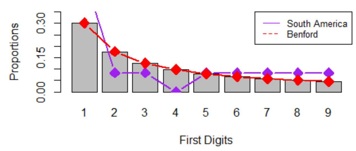
(c) Distribution of first digits cumulative active cases: Australia



(d) Distribution of first digits cumulative active cases: Europe



(e) Distribution of first digits cumulative active cases: North America



(f) Distribution of first digits cumulative active cases: South America

Fig. 5: First digits distribution between Benford and World continental COVID-19 active cases

but rather due to its smaller sample size (with limited number of data points of 6 for each variable considered). Per the results of this study, COVID-19 reported data of individual countries constituting the Australia/Oceania continent including Australia, New Zealand, Papua New Guinea, French Polynesia, Fiji and New Caledonia, were inspected to ascertain the shocking revelations of this study. A careful examination of the cumulative confirmed cases, cumulative deaths, cumulative recovered cases and cumulative active cases has it that except for Australia and New Zealand, which had five variables with four or more digits, all the other countries had three or fewer digits as of August 4, 2020, which is not ideal data to achieve Benford's conformity.

4 Conclusions and Recommendations

In this study, Newcomb Benford's Law for first digits with focus on the [48] d^*, [49] Statistic (α^*), [50] Statistic (ω^*) and Kolmogorov-Smirnov test to detect inconsistencies in daily COVID-19 reportage cases and deaths during the first wave of the pandemic of all world continents were considered. Likewise, histograms depicting density distributions of first digits were built and plot together to Newcomb Benford's density distribution of first digits. We used the deviation from the Newcomb–Benford law of anomalous numbers as a proxy for data accuracy. Although the applicability of Benford's Law to assess world continental COVID-19 data has been established by other researchers, this study however provided a confirmatory test to determine its validity. In this study, approximately one-sixth of the world's continents, we document some evidence of deviations from the Newcomb Benford's Law which is subject to further investigation or data verification. In general, the results presented in this study indicate that there are possible no or minimal indications of world COVID-19 data reportage alteration during the first wave of the pandemic. Except for the Australia/Oceania continent where there is strong evidence that the data structure was not a good fit for Newcomb Benford's Law analysis at the time of data acquisition (with less than four digits data counts for a host of related variables during the first wave of the COVID-19 pandemic) as noted by [51] that small samples may have insufficient power to meaningfully detect or confirm conformance with Benford's law. In general, the study found that world continental COVID-19 data reportage during the first wave of the COVID-19 pandemic was dependable though these findings are subject to further scrutiny. However, it is necessary to carry out further studies on the individual countries that constitute the Australia/Oceania continent in order to confirm possible anomalies/manipulation as suggested by Benford's Law in this study or otherwise as recommended by [54].

The continental results presented in this study are in perfect agreement with other country-specific results by other researchers using Newcomb Benford's law. For example, in Indonesia (see [56]), in China (see [55]), in European countries like Bulgaria, Croatia, Lithuania and Romania (see [53]), in South American countries like Mexico and Brazil (see [52]). This study has established that the transparency and integrity of COVID-19 data reporting across all world continents were consistently maintained during the first wave of the COVID-19 pandemic when the world was in search of a reliable vaccine to combat the pandemic. For anomaly detection, the Benford test can be a valuable statistical technique. While the data from other continents exhibited conformity to Benford's Law, it is important to note that Benford's Law is not infallible and should be used as a complementary tool rather than the sole determinant of data reliability. There may be legitimate reasons for data to deviate from Benford's Law, such as variations in testing capabilities, reporting protocols, and different levels of transparency among countries. Therefore, to assess the validity and reliability of a dataset, researchers and analysts should start with the examination of the distribution of leading digits using the Newcomb Benford test. When a dataset complies with Benford's Law, then it can be confidently used, while deviations require further investigation. To improve the accuracy and reliability of anomaly detection processes, other anomaly detection methods and domain-specific knowledge must be combined with the Benford test using at least the first two digits. Thus, to establish a more definitive conclusion when the Benford test provided inconclusive results, additional analyses and complementary anomaly detection methods such as density-based techniques including local outlier factor and isolation forests adopted by [66] are recommended.

Declarations

Competing interests: Authors have no conflict of interest
Authors' contributions: Edmund F. Agyemang: conceptualization, methodology, writing original draft, visualization, reviewing, editing, project administration, methodology, formal analysis, supervision, resources, investigation, and validation. Joseph A. Mensah: writing original draft, project administration, resources, visualization, supervision, investigation, reviewing and validation. Eric Nyarko: writing original draft, reviewing, resources, editing, project administration, and validation. Ezekiel N.N. Nortey: Reviewing, editing, methodology, project administration, resources, investigation, supervision, and validation.
Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
Availability of data and materials: The data used to support the findings of this study are available from the corresponding author upon reasonable request.
Acknowledgments: The authors thank the anonymous reviewers and editors whose insightful comments helped enriched the work. The first and corresponding author acknowledges the enormous support of the University of Texas Rio Grande Valley Presidential Research Fellowship fund.

References

[1] H. R. Güner, I. Hasanoğlu, F. Aktaş, COVID-19: Prevention and control measures in community, *Turkish Journal of medical sciences*, **50(9)**(2020), 571–577.

[2] L. Kuznetsova, COVID-19: The world community expects the World Health Organization to play a stronger leadership and coordination role in pandemics control, *Frontiers in Public Health*, **8**(2020), 470.

[3] M. Jit, A. Ananthakrishnan, M. McKee, O.J. Wouters, P. Beutels, Y. Teerawattananon, Multi-country collaboration in responding to global infectious disease threats: lessons for Europe from the COVID-19 pandemic, *The Lancet Regional Health-Europe*, **9**(2021), 100221.

[4] C. Su, K. Dai, S. Ullah, Z. Andlib, COVID-19 pandemic and unemployment dynamics in European economies, *Economic Research-Ekonomska Istraφzivanja*, ***35(1)***(2022), 1752 − −1764.*X.Shi, G.Y.Brasseur, TheresponseinairqualitytothereductionofChineseeconomicactivitiesduringtheCOVID* − 19*outbreak*, Geophysical Research Letters, ***47(11)***(2020), *e*2020*GL*088070.

[5] *E. Nyarko, E.F. Agyemang, D. Arku, COVID-19 and online teaching in higher education: A discrete choice experiment,* Model Assisted Statistics and Applications, ***18(2)****(2023), 115–123.*

[6] *K. Mukhtar, K. Javed, M. Arooj, A. Sethi, Advantages, Limitations and Recommendations for online learning during COVID-19 pandemic era,* Pakistan journal of medical sciences, ***36****(2020), S27.*

[7] *O.B. Adedoyin, E. Soykan, Covid-19 pandemic and online learning: the challenges and opportunities,* Interactive learning environments, ***31(2)****(2023), 863–875.*

[8] *N.A. Bakar, S. Rosbi, Effect of Coronavirus disease (COVID-19) to tourism industry,* International Journal of Advanced Engineering Research and Science, ***7(4)****(2020), 189–193.*

[9] *J. Abbas, R. Mubeen, P.T. Iorember, S. Raza, G. Mamirkulova, Exploring the impact of COVID-19 on tourism: transformational potential and implications for a sustainable recovery of the travel and leisure industry,* Current Research in Behavioral Sciences, ***2****(2021), 100033.*

[10] *D.C. Wu, C. Cao, W. Liu, J.L. Chen, Impact of domestic tourism on economy under COVID-19: The perspective of tourism satellite accounts,* Annals of Tourism Research Empirical Insights, ***3(2)****(2022), 100055.*

[11] *J.A. Duro, A. Osório, A. Perez-Laborda, Competition for domestic tourism in the COVID-19 pandemic: A characterization using a contest model,* Tourism economics, ***29(2)****(2023), 378–391.*

[12] *O. Okoi, T. Bwawa, How health inequality affect responses to the COVID-19 pandemic in Sub-Saharan Africa,* World development, ***135****(2020), 105067.*

[13] *D. Mensah, E.F. Agyemang, C. Gewa, Understanding Women, Infant and Children Feeding Decisions in Emergencies: The Case of COVID-19 and the Formula Shortage,* Journal of Health, Medicine and Nursing, ***118****(2024), 49–64.*

[14] *C. Bambra, R. Riordan, J. Ford, F. Matthews, The COVID-19 pandemic and health inequalities,* J Epidemiol Community Health, ***74(11)****(2020), 964–968.*

[15] *S.A. Mein, COVID-19 and health disparities: the reality of "the Great Equalizer",* Journal of general internal medicine, ***35(8)****(2020), 2439–2440.*

[16] *D. Fanelli, F. Piazza, Analysis and forecast of COVID-19 spreading in China, Italy and France,* Chaos, Solitons & Fractals, ***134*** *(2020).*

[17] *P. Matjaφz, G.Miksić, N.Slavinec, M.Stoφzer, Forecastingcovid* − 19, Frontiers in Physics, (2020), 127.

[18] *Z. Abdelhafid, H. Fouzi, D. Abdelkader, S. Ying, Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study,* Chaos, Solitons and Fractals, ***140*** *(2020),110121.*

[19] *J.L. Christopher, Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months,* MedRxiv,*(2020).*

[20] *C. Matteo, D.T. Jessica, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, P. Piontti, The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak,* Science, ***368(6489)*** *(2020), 395–400.*

[21] *M.A.Moret, and J.F.F. Mendes, A transnational and transregional study of the impact and effectiveness of social distancing for COVID-19 mitigation,* Entropy, ***23(11)****(2021) , 1530.*

[22] *D.E. Natalie, J.W. Hogan, M.E. Schnitzer, Covid-19 vaccine effectiveness and the test-negative design,* New England Journal of Medicine, ***385(15)*** *(2021), 1431–1433.*

[23] *S. Joffe, and A. Babiker, S.S. Ellenberg, A. Fix, M.R. Griffin, S. Hunsberger, J. Kalil, M.M. Levine, M.W. Makgoba, R.H. Moore, Data and safety monitoring of COVID-19 vaccine clinical trials,* The Journal of infectious diseases, ***224(121)****(2021), 1995–2000.*

[24] *T. Yamaguchi, M. Iwagami, C.Ishiguro, D. Fujii, N. Yamamoto, M. Narisawa, T. Tsuboi, H. Umeda, N. Kinoshita, T. Iguchi, Safety monitoring of COVID-19 vaccines in Japan,* The Lancet Regional Health-Western Pacific, ***23****(2022).*

[25] *M. Tenforde, Effectiveness of Pfizer-BioNTech and Moderna vaccines against COVID-19 among hospitalized adults aged >= 65 years—United States, January–March 2021,* MMWR. Morbidity and mortality weekly report, ***70*** *(2021).*

[26] *V.H. Moreau, Inconsistencies in countries COVID-19 data revealed by Benford's law,* Model Assisted Statistics and Applications, ***16(1)****(2021), 73–79.*

[27] *A. Karlinsky, D. Kobak, Tracking excess mortality across countries during the COVID-19 pandemic with the World Mortality Dataset,* Elife, ***10****(2021), e69336.*

[28] *M.F. Prado, B.B. Antunes, L.L Bastos,I.T. Peres, A.A.B Silva, L.F. Dantas, F.A Baião,P. Maçaira, S. Hamacher, F.A Bozza, Analysis of COVID-19 under-reporting in Brazil,* Journal, ***32****(2020), 224–228.*

[29] E. Kupek, How many more? Under-reporting of the COVID-19 deaths in Brazil in 2020, Tropical Medicine & International Health, ***26(9)****(2021), 1019–1028.*

[30] P. Wang,T. Hu, S. Bao, H. Liu, X. Zhu, Exploring the impact of under-reporting cases on the spatiotemporal distributions of COVID-19: The case of Hubei, China, arXiv preprint arXiv:2011.04892*, (2020).*

[31] N. Thenon, M. Peyre, M. Huc, A. Touré, F. Roger, S. Mangiarotti, COVID-19 in Africa: Underreporting, demographic effect, chaotic dynamics, and mitigation strategy impact, PLoS Neglected Tropical Diseases*,* ***16(9)****(2022), e0010735.*

[32] M.R. Tovani-Palone, F. Garoli, P.A. Shah, M.A. Kamal, F.A. Nawaz, Underreporting of cases during the COVID-19 pandemic: a worrying warning for Africa, Journal of Contemp Stud Epidemiol Public Health*,* ***3(1)****(2022).*

[33] P. Nguimkeu, S. Tadadjeu, Why is the number of COVID-19 cases lower than expected in Sub-Saharan Africa? A cross-sectional analysis of the role of demographic and geographic factors, World Development*,* ***138****(2021), 105251.*

[34] H. Lau, T. Khosrawipour, P. Kocbach, H. Ichii, J. Bania, V. Khosrawipour, Evaluating the massive underreporting and undertesting of COVID-19 cases in multiple global epicenters, Pulmonology*,* ***27(2)****(2021), 110–115.*

[35] S. Krantz, A.S. Rao, Arni, Level of underreporting including underdiagnosis before the first peak of COVID-19 in various countries: Preliminary retrospective results based on wavelets and deterministic modeling, Infection Control & Hospital Epidemiology*,* ***41(7)****(2020), 857–859.*

[36] R.F. Reis, B. de Melo Quintela, J. de Oliveira Campos, J.M. Gomes, B.M. Rocha, M. Lobosco, R.W. Dos Santos, Characterization of the COVID-19 pandemic and the impact of uncertainties, mitigation strategies, and underreporting of cases in South Korea, Italy, and Brazil, Chaos, Solitons & Fractals*,* ***136****(2020), 109888.*

[37] N. Farhadi, Can we rely on COVID-19 data? An assessment of data from over 200 countries worldwide, Science progress*,* ***104(2)****(2021).*

[38] C. da Silva Azevedo, R.F. Gonçalves, V.L. Gava, M. de Mesquita Spinola, A Benford's Law based methodology for fraud detection in social welfare programs: Bolsa Familia analysis, Physica A: Statistical Mechanics and its Applications*,* ***567****(2021), 125626.*

[39] C. Goh, Applying visual analytics to fraud detection using Benford's law, Journal of Corporate Accounting & Finance*,* ***31(4)****(2020),202–208.*

[40] D. Crisan, A. Irimia, D. Gota, L. Miclea, A. Puscasiu, O. Stan, H. Valean, Analyzing Benford's Law's Powerful Applications in Image Forensics, Applied Sciences*,* ***11(23)****(2021), 11482.*

[41] S. Tunmibi, W. Olatokun, Application of digits based test to analyse presidential election data in Nigeria, Commonwealth & Comparative Politics*,* ***59(1)****(2021), 1–24.*

[42] E.F. Agyemang, E.N.N. Nortey, R. Minkah, K. Asah-Asante, The unfolding mystery of the numbers: First and second digits based comparative tests and its application to Ghana's elections, Model Assisted Statistics and Applications*,* ***18(2)****(2023), 183–192.*

[43] E.F. Agyemang, E.N.N. Nortey, R. Minkah, K. Asah-Asante, Baseline Comparative Analysis and Review of Election Forensics: Application to Ghana's 2012 and 2020 Presidential Elections, Heliyon*, (2023).*

[44] E.N.N. Nortey, E.F. Agyemang, R. Minkah, K. Asah-Asante, Bayesian estimation of presidential elections in ghana: A validation approach, African Journal of Applied Statistics*,* ***9(1)****(2022), 1297–1317.*

[45] A. Kilani, Authoritarian regimes' propensity to manipulate Covid-19 data: a statistical analysis using Benford's Law, Commonwealth & Comparative Politics*,* ***59(3)****(2021), 319–333.*

*[46] M.R. Fewster, A simple explanation of Benford's Law,*The American Statistician*,****63(1)****(2009), 26–32.*

[47] L. Liberti, C. Lavor, N. Maculan, A. Mucherino, Euclidean distance geometry and applications, SIAM review*,* ***56(1)****(2014), 3–69.*

[48] T. Cho, K. W.K. Gaines, J. Brian, Breaking the (Benford) law: Statistical fraud detection in campaign finance, The American statistician*,****61(31)****(2007), 218–223.*

[49] L.M. Leemis, B.W. Schmeiser, D.L. Evans, Survival distributions satisfying Benford's law, The American Statistician*,* ***54(4)****(2000), 236–241.*

[50] W.K. Tam Cho, B.J. Gaines, Statistical fraud detection in campaign finance. Breaking the (Benford) law, Am. Stat*,* ***61****(2007), 218–223.*

[51] W. Goodman, The promises and pitfalls of Benford's law, Significance*,* ***13(3)****(2016), 38–41.*

[52] A.P. Kennedy, S.C.P. Yam, On the authenticity of COVID-19 case figures, PloS one*,* ***15(12)****(2020), e0243123.*

[53] P. Kolias, Applying Benford's law to COVID-19 data: the case of the European Union, Journal of Public Health*,* ***44(2)****(2022), e221–e226.*

[54] R. Isea, How Valid are the Reported Cases of People Infected with Covid-19 in the World?, International Journal of Coronaviruses*,* ***1****(2020), 53–56.*

[55] C. Koch, K. Okamura, Benford's law and COVID-19 reporting, Economics letters*,* ***196*** *(2020), 109573.*

[56] W.K. Wong, F.H. Juwono, W.N. Loh, I.Y. Ngu, Newcomb-Benford law analysis on COVID-19 daily infection cases and deaths in Indonesia and Malaysia, Heritage and Sustainable Development*,* ***3(2)****(2021), 102–110.*

[57] W.M. Goodman, Reality checks for a distributional assumption: The case of "Benford's Law", Joint Statistical Meeting–Business and Economic Statistics Section. Citeseer*, (2013), 2789–2803.*

[58] A.E. Kossovsky, On the mistaken use of the chi-square test in Benford's law, Stats*,* ***4(2)****(2021), 419–453.*

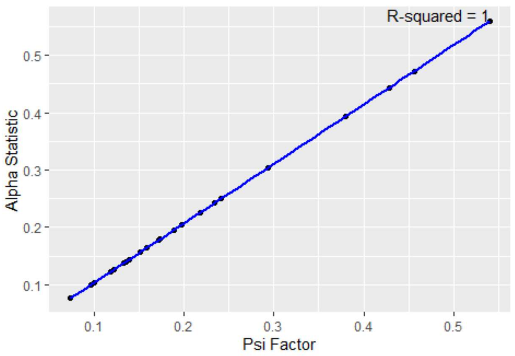
[59] T.P. Hill, The significant-digit phenomenon, The American Mathematical Monthly*,* ***102(4)****(1995), 322–327.*

[60] J. Hedderich, L. Sachs, Angewandte Statistik, Journal*, (2016).*

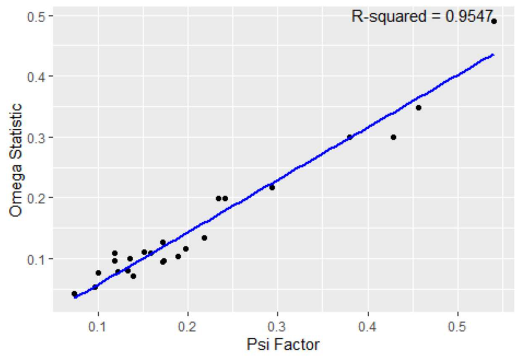
[61] J. Massey, J. Frank, The Kolmogorov-Smirnov test for goodness of fit, Journal of the American statistical Association*,* ***46(253)****(1951), 68–78.*

[62] V.S. Balashov, Y. Yan, X. Zhu, Using the Newcomb–Benford law to study the association between a country's COVID-19 reporting accuracy and its development, Scientific reports, ***11(1)****(2021), 22914.*
[63] D.E. Giles, Benford's law and naturally occurring prices in certain ebaY auctions, Applied Economics Letters, ***14(3)****(2007), 157–161.*
[64] M.A. Stephens, Use of the Kolmogorov–Smirnov, Cramer–Von Mises and related statistics without extensive tables, Journal of the Royal Statistical Society: Series B (Methodological), ***32(1)****(1970), 115–122.*
[65] N.H. Kuiper, Tests concerning random points on a circle, Nederl. Akad. Wetensch. Proc. Ser. A, ***63(1)****(1960), 38–47.*
[66] E.F. Agyemang, Anomaly detection using unsupervised machine learning algorithms: A simulation study, Scientific African, ***26****(2024), e02386.*

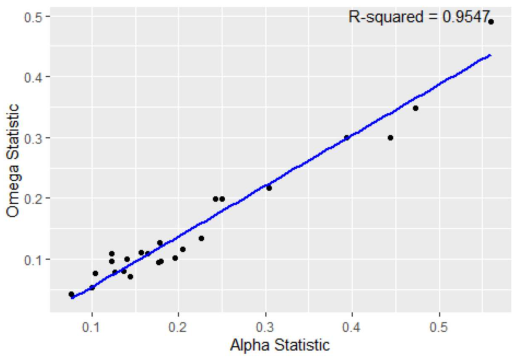
5 Appendix



(a) Correlation between d^* and α^*: Whole World



(b) Correlation between d^* and ω^*: Whole World



(c) Correlation between α^* and ω^*: Whole World

Fig. 6: Correlation between d^*, α^* and ω^* computed by fitting of the first digits distribution of studied world continental COVID-19 data into the Benford's distribution in Equation 8.