ANALYSIS OF COVID-19 CONTAMINATION AND DEATHS CASES IN BRAZIL ACCORDING TO THE NEWCOMB-BENFORD LAW

Carlos Roberto Souza CARMO[1]
Fernando de Lima CANEPPELE[2]
Fábio Caixeta NUNES[2]

- ABSTRACT: The use of the Newcomb-Benford Law in assessing the quality of health and / or epidemiological information systems can allow relevant decisions to be made to improve these systems. In this context, this research aimed to carry out an assessment of the conformity of the information regarding the number of cases of contamination and deaths by COVID-19 in Brazil according to the Newcomb-Benford Law, from the moment of the occurrence of the first case of the disease and from the first death by COVID-19 in the country until the month of September 2020. With the aid of descriptive statistics and the use of metrics related to the Z test and the mean absolute deviation it was possible to observe that, both from a national and longitudinal perspective as for the transversal-state perspective, the quantitative data referring to the cases of contamination by the coronavirus and the deaths that occurred as a result of COVID-19 did not present the expected behavior according to the Newcomb-Benford Law. Due to the lack of conformity in relation to the Newcomb-Benford Law, it is suspected that some level of conformity specific to this type of data has occurred, in the Brazilian context, since there are already studies that suggest the existence of proper levels of conformity for certain types of data.
- KEYWORDS: Epidemiology; conformity; quantitative methods applied.

1 Introduction

In December 2019, the outbreak of a new infectious disease caused by the coronavirus, COVID-19, began in Wuhan city, Hubei province, China (IDROVO and MANRIQUE-HERNÁNDEZ, 2020). On March 11, 2020, the World Health Organization (WHO) declared that the outbreak of COVID-19 will become a global pandemic (PAN and ZHANG, 2020). By the end of March 2020, COVID-19 had already spread around the world, contaminating 850,000 more people and causing more than 40,000 deaths (IDROVO and MANRIQUE-HERNÁNDEZ, 2020).

In these circumstances, several nations had already been preparing to deal with this serious threat to world health. However, in some cases, this preparation took a little longer to get underway. By June 2020, COVID-19 had reached 213 countries, generated more than 9 million confirmed cases, and caused more than 470,000 deaths around the world

[1] Universidade Federal de Uberlândia – UFU, Departamento de Ciências Contábeis, CEP: 38400-902, Uberlândia, MG, Brasil. Email: *carlosjj2004@hotmail.com*
[2] Universidade de São Paulo – USP – Departamento de Engenharia de Biosistemas, CEP 13635-900, Pirassununga, SP, Brasil. Email: *caneppele@usp.br*, *fabiocaixeta@usp.br*

(PAN and ZHANG, 2020). Amidst the speed of contamination imposed by the coronavirus, the increasing number of deaths generated from COVID-19, and the trials, hit and miss on a global scale, a variety of information has been produced and disseminated.

In crises such as the pandemic of COVID-19, where the generation of data occurs almost instantaneously from isolated sources and/or of varying natures, it becomes imperative that the entities involved coordinate and collaborate quickly and effectively to combat the pandemic (PAN and ZHANG, 2020).

The implementation of epidemiological surveillance systems is vital in combating the spread and consequences of the disease. Since one of its most critical functions focuses on providing information of sufficient quality to support decision making, which must be evidence-based. However, producing accurate and reliable information throughout emergencies such as the pandemic of COVID-19 is not a simple task (IDROVO and MANRIQUE-HERNÁNDEZ, 2020).

Because information systems are particularly important during public health emergencies such as the COVID-19 pandemic, they need to be constantly evaluated and adjusted. In such a way that a continuous and effective flow of information is maintained. In this sense, the use of the Newcomb-Benford Law (1881, 1938) can be a useful tool due to its fast and objective approach. Furthermore, its ease of application in analyzing the reliability of data of an empirical-quantitative nature.

As noted by Idrovo and Manrique-Hernández (2020), the evaluation provided by the use of the Newcomb-Benford Law (1881, 1938) in the process of validating the quality of health information systems can enable relevant decision making for the improvement of epidemiological surveillance systems aimed at supporting and targeting actions to combat COVID-19.

In this context, this research aimed to evaluate the conformity of the information regarding the number of cases of contamination and deaths by COVID-19 in Brazil, according to the Newcomb-Benford Law (1881, 1938), from the moment of the first case of the disease and the first reported death in the country until September 2020.

In this sense, this scientific research process was conducted based on the following guiding question: according to the Newcomb-Benford Law (1881, 1938), what is the degree of conformity of the information reported in Brazil regarding the cases of contamination (new and accumulated) and deaths (new and accumulated) due to COVID-19 during the period between 02/25/2020 and 09/24/2020?

For this purpose, initially, the theoretical background about the theme related to the Newcomb-Benford Law (1881, 1938) was performed. Its areas of application were identified, as well as the results of studies related to its specific application in the field of public health surveillance, with special attention to the evaluation of the quality of information about the pandemic of COVID-19, as can be seen in the second section of this article.

Next, the database containing the information about the cases of contamination and deaths that occurred as a result of COVID-19 between 02/25/2020 and 09/24/2020 in Brazil was investigated and composed, and the respective data analysis methodology was identified, as described in the third section of this research.

From the theoretical platform constituted for this study and the respective research sample, the data analysis was performed, and the results arising from the use of the

Newcomb-Benford Law (1881, 1938) were presented, as reported in the fourth section of this paper.

Finally, in the fifth section of this research, the final considerations about the whole scientific investigation and suggestions for its continuity were made.

2 Theoretical background

In reviewing a book of logarithmic tables in 1881, Simon Newcomb observed that the tables with smaller numbers were more worn than those with larger numbers, and from this observation, Newcomb (1881) concluded that "the law of probability of the occurrence of numbers is such that all the mantissas of their logarithms are equally probable."

After 57 years, Benford (1938) concluded such as that first proposed by Newcomb (1881) by identifying a data set composed of more than 20,000 observations referring to a diversity of real phenomena whose behavior followed the probability law proposed by Newcomb (1881).

This law has since come to be recognized as the Newcomb-Benford Law, henceforth just referred to as the NB Law, and has been the subject of many studies over the years.

The NB Law assumes that when looking at various sets of data representing the occurrence of real-life events or mathematical tables, among other publications, it can be seen that the most significant digits of the totalizing numbers of these occurrences have a higher probability of being started by digits of smaller value, obeying a very particular logarithmic distribution (LEE *et al*., 2020), as described in Equation 1 (BENFORD, 1938).

$$F_a = log\left(\frac{a+1}{a}\right) \quad (1)$$

Thus, according to NB Law, the occurrence of naturally generated significant digits tends to follow the frequency distribution described in Table 1 (DRUICĂ *et al.*, 2018).

Table 1 - Frequency distribution of significant digits according to NB Law

Digit	1th place	2nd place	3rd place
1	0.30103	0.11389	0.10138
2	0.17609	0.10882	0.10097
3	0.12494	0.10433	0.10057
4	0.09691	0.10433	0.10018
5	0.07918	0.09668	0.09979
6	0.06695	0.09337	0.09940
7	0.05799	0.09035	0.09902
8	0.05115	0.08757	0.09864
9	0.04576	0.08500	0.09827
0	-	0.11968	0.10178

Source: Elaborated by the authors from Druică *et al*. (2018).

The research on the applications of the NB Law has already addressed study objects of the most varied natures: genome data; half-life of unstable nuclei; toxic emissions data; tax auditing; accounting; electoral data; stock markets; regression coefficients; inflation data; religions; birth data; rivers; first letter words; decay rates of elementary particles; astrophysical measurements, among others (LEE *et al.*, 2020).

According to Manrique-Hernández *et al.* (2020), the first study related to the application of the NB Law in the field of public health surveillance came from the research of Gómez-Camponovo *et al.* (2016). Also, according to Manrique-Hernández *et al.* (2020), in the field of epidemics the first scientific works of this nature were the research of Manrique-Hernández *et al*. (2017) and Idrovo and Manrique-Hernández (2020).

Gómez-Camponovo *et al.* (2016) evaluated the performance of Paraguay's epidemiological surveillance system between 2009 and 2011 by comparing the distribution of reported dengue cases with the distribution predicted according to the NB Law, for first and second digits of records, globally and separated by region, season, population density, among other stratifications.

Gómez-Camponovo *et al.* (2016) concluded that the Paraguayan epidemiological surveillance system worked better in non-epidemic periods and in states with better accommodation structures for the bovine herd. With regard to the application of the NB Law, it was highlighted that its use and practicality allowed the rapid assessment of data quality and the sensitivity of epidemiological surveillance systems (GÓMEZ-CAMPONOVO *et al*., 2016)

Using the NB Law to evaluate the overall performance of epidemiological surveillance systems about cases of contamination by Zika virus, with the epidemic still in progress, Manrique-Hernández *et al*. (2017) call attention to the fact that experiences with this type of analysis have shown promise. And, in the context of this issue, the application of the NB Law proved to be able to quickly identify places where surveillance could be improved (MANRIQUE-HERNÁNDEZ *et al*., 2017).

As Idrovo and Manrique-Hernández (2020) point out, reliable epidemic information systems are essential during public health emergencies, as is the case of the COVID-19 pandemic. In this context, the adjustment of NB Law to empirical data can be useful to objectively and quickly assess the performance of surveillance systems during epidemics and pandemics (IDROVO and MANRIQUE-HERNÁNDEZ, 2020).

Based on information available in WHO situational reports, Idrovo and Manrique-Hernández (2020) applied NB Law to data regarding the number of confirmed cases, suspected cases and deaths in the past 24 hours, and cumulative confirmed cases and deaths, occurring in each province, region and city in China over the period from January 21 to March 15, 2020, to assess the quality of Chinese surveillance data about COVID-19.

The results identified by the research of Idrovo and Manrique-Hernández (2020) concluded that the NB Law provided a rapid assessment that suggested that China's epidemiological surveillance system had good quality data, considering that it was the country of origin of the pandemic.

In another study, Manrique-Hernández *et al.* (2020) used the NB Law to evaluate the performance of Colombia's epidemiological surveillance system over the first 50 days of the COVID-19 pandemic. In that investigation, the quality of data according to the NB Law and the timeliness of information, measured from the difference in dates between the

generation of data generated at the National Health Institute of Colombia and the data collected in the WHO situational report, were evaluated (MANRIQUE-HERNÁNDEZ *et al*., 2020).

As the main results, Manrique-Hernández *et al.* (2020) reported that the NB Law was observed in most of the days evaluated and also that the timeliness of the information depended basically on the difference between the closing times of the information by the National Health Institute of Colombia and the compilation of data by the WHO. Overall, the study by Manrique-Hernández *et al.* (2020) suggests that the information provided by the Colombian epidemiological surveillance system followed the NB Law, evidencing data quality.

Lee *et al*. (2020) proposed an epidemic growth model that could capture the effects of interventions by various countries to flatten the growth curve of COVID-19 infections and could relate the model and the NB Law to assess the existence of fraud in the number of deaths self-reported by these countries.

Among other results, the analysis by Lee *et al.* (2020) identified a diversity of growth curve profiles of COVID-19 infections, with growth deceleration parameter estimates ranging from 0.549 to 0.999 and conformance to the NB Law, except for information from Japan.

Lee *et al.* (2020) assessed that one of the main consequences of the evidence collected by their study was that researchers may feel discouraged to conduct researches based on Japan's self-reported information.

Given the above, the use of the NB Law in the process of evaluating the quality of information produced from the COVID-19 pandemic has shown promise, proving to be an agile, objective, and easily understood tool from an empirical point of view, although this methodology does not allow the identification of the causes of inconsistencies in the epidemiological information systems analyzed.

3 Material and methods

Initially, the data in CSV format (comma-separated values files) available in the case panel of coronavirus disease 2019 (COVID-19) prepared by the Brazilian Ministry of Health (BRASIL, 2020) were identified. However, preliminary analysis of the data from the Ministry of Health of Brazil revealed that information on COVID-19 cases only began to be detailed by municipalities as of 03/27/2020, although the first case of contamination was reported by the health secretary of the state of São Paulo on 02/25/2020 and the first death occurred on 03/17/2020, also in the state of São Paulo.

Next, the repository of public data made available by the Brasil.IO (2020) website was searched. This website compiles, stores, and reports daily data regarding confirmed cases and deaths obtained directly from the bulletins of the State Health Secretariats (SES) in Brazil, among other information. The data in question were downloaded (file *<caso_full.csv>)* in CSV format and treated using electronic calculation spreadsheets.

After tabulating the data made available by the Brasil.IO website (2020) at the municipal level, according to its totalization by each of the 27 Brazilian federative units (26 states and the Federal District), the conference of the respective totals was performed with the information made available by the Brazilian Ministry of Health (BRASIL, 2020), but at the state level until 03/26/2020, as already informed. It was observed that, despite

the absence of initial data at the municipal level, the information made available by the Brazilian Ministry of Health (BRASIL, 2020) corresponded at the state level with the data informed by the SES and made available on the Brasil.IO website (2020).

The use of two databases from different sources (one government source and one independent source) is justified by the fact that, on June 5th, 2020, the Brazilian federal government deliberately decided to no longer publish the accumulated numbers of cases and deaths resulting from the pandemic of COVID-19 in the country. Both in its daily bulletins and in the website created to deal with the subject (BARIFOUSE, 2020). This decision was criticized by several sectors in Brazil (politicians, judiciary, scientists, etc.) and was also echoed by the WHO director of emergencies (BARIFOUSE, 2020).

Although the Brazilian government has reversed its decision, several sites were identified that make such information available on their initiative. Among them, Brasil.IO was highlighted, whose data are used in the composition of the MonitoraCovide-19 site, created by the Oswaldo Cruz Foundation (BARIFOUSE, 2020).

After comparing and validating the data from those two distinct sources, one in counter examination to the other, the file with the data used in this study was finalized. Thus, 213 records were identified regarding the daily totals of new cases of contamination by COVID-19, and also, with the total of daily accumulated cases, according to the SES of the 27 states of Brazil, during the period between 02/25/2020 and 09/24/2020. We also identified information regarding 192 records with the daily totals of new deaths from the COVID-19 and total daily accumulated deaths, informed by the SES of the 27 states of Brazil, during 03/17/2020 and 09/24/2020. It is worth pointing out that the first death only occurred on 03/17/2020, in the state of São Paulo, despite the pandemic had arrived in Brazil on 02/25/2020.

Initially, data analysis was performed from a longitudinal perspective at the national level, in which the first digits of the daily totals of cases (new and accumulated, from 02/25/2020 to 09/24/2020, therefore, 213 days) and the daily total of deaths (new and accumulated, from 03/17/2020 to 09/24/2020, therefore, 192), for all of Brazil, were evaluated.

In this step, the Z-test was used to evaluate the hypothesis related to the existence of statistically significant differences between the observed relative frequencies (%) for the analyzed data *(Fro)* and the expected relative frequency (%) according to NB Law *(Fre)*.

Considering that the Z test tends to be influenced by the size of the sample analyzed, becoming more rigorous for more numerous samples (COSTA *et al.*, 2012), the mean absolute deviation (MAD) was used in a complementary way to corroborate the results obtained from the use of the Z test.

The MAD consists of the average absolute deviation calculated by dividing the sum of the absolute deviations by the number of significant digits (9 for the first position and 10 for the second), as described by Equation 2.

$$MAD = \frac{\sum_{i=1}^{k} |Fro_i - Fre_i|}{k} \quad (2)$$

If, on the one hand, the MAD is not influenced by the size of the sample analyzed as happens with the Z test, on the other hand, it does not allow calculating limits or ranges of

values within which it can be considered significant or not (COSTA *et al.*, 2012). However, Nigrini (2012) provides a parameter from which it is possible to establish degrees of conformity with the NB Law for MAD analysis, as presented in Table 2.

Table 2 - Critical values for analyzing the mean absolute deviation from NB Law

Digit	Full conformity	Acceptable conformity	Marginal conformity	Non-conformity
First	0.000 to 0.006	0.006 to 0.012	0.012 to 0.015	Higher than 0.015
Second	0.000 to 0.008	0.008 to 0.010	0.010 to 0.012	Higher than 0.012
First two digits	0.000 to 0.012	0.012 to 0.018	0.018 to 0.022	Higher than 0.022

Source: prepared by the author based on Nigrini (2012).

After the longitudinal analysis of the data at the national level, a new round of analysis was carried out. However, this time, in a cross-sectional and state level, that is, the first digits of the totals reported for each Brazilian state and the Federal District were analyzed for 213 days (from 02/25/2020 to 09/24/2020) for new and accumulated cases of COVID-19. Besides, we analyzed the first digits of the totals reported for each Brazilian state and the federal district over 192 days (03/17/2020 to 09/24/2020) for new and cumulative deaths due to the pandemic of COVID-19 in Brazil.

Given the results of the application of the Z-test in the analysis of longitudinal data at the national level, together with the results of the application of the MAD, in this new analytical step, only the MAD and the conformity analysis proposed by Nigrini (2012) were used as parameters. To assess the degree of conformity between the relative frequencies (%) observed for the data analyzed (*Fro*) and the expected relative frequency (%) according to the NB Law (*Fre*), whose results were summarized by presenting descriptive statistics relating to the mean, median, standard deviation, coefficient of variation, minimum and maximum values observed, and also the amplitude of those data.

The justification for this procedure was based on the cost factor compared to the analytical benefits generated using the Z-test, i.e., since the first analytical step (longitudinal data analysis at the national level) the MAD fully corroborated the Z-test results. In this case, it was admitted fact that MAD represents a more operational analytical methodology when considering the number of treatments and calculations required for the application of the Z test. Based on the conformity analysis proposed by Nigrini (2012), MAD evaluates the general conformity of a whole series of digits i, with i=1, ..., k, for each federative unit, while the Z test would demand a digit-by-digit frequency analysis for each federative unit.

When considering the object of study of this scientific investigation, the respective analytical methods, and the nature of the data analyzed, this study can be classified as scientific research of an empirical nature whose data analysis was based on applied quantitative methods.

4 Data analysis and results

The analysis of the data in longitudinal perspective about the first digits of the totals of new cases of COVID-19 reported daily in Brazil, throughout the period from

02/25/2020 to 09/24/2020, revealed that only the observed frequencies for digits 1, 2, 5, and 6 did not present statistically significant differences, at a significance level of 5% (Z-statistic with *p-value* > 0.05), concerning the expected relative frequency according to the NB Law, as shown by the information summarized in Table 3.

Table 3 - Analysis of the first digit of daily totals of new cases

1st digit	Relative Frequency		Proportion Test	
	expected (*Fre*)	observed (*Fro*)	Z-Statistic	*p-value*
1	30.10%	24.27%	-1.75	0.081
2	17.61%	19.42%	0.59	0.555
3	12.49%	18.93%	2.69	0.007
4	9.69%	20.87%	5.31	0.000
5	7.92%	8.25%	0.05	0.962
6	6.69%	4.37%	-1.19	0.233
7	5.80%	1.94%	-2.22	0.026
8	5.12%	0.49%	-2.86	0.004
9	4.58%	1.46%	-1.98	0.048

Source: Elaborated by the authors based on the survey data.

This result was corroborated by the analysis performed from the MAD calculation, whose conformity analysis proposed by Nigrini (2012) signaled an overall non-conformity concerning the NB Law, in which the observed MAD was greater than the critical MAD for non-conformity (0.0439 or 4.39% > 0.015 or 1.50%).

Regarding the analysis of the first digits of the cumulative totals of cases of contamination by COVID-19 in Brazil, the totals with first digit equal to 1, 2, 5, 6, 8, and 9 did not show statistically significant differences at a significance level of 5% (Z-statistic with *p-value* > 0.05), about the expected relative frequency according to the NB Law, as can be seen in Table 4.

Table 4 - Analysis of the first digit of daily cumulative case totals

1st digit	Relative Frequency		Proportion Test	
	expected (*Fre*)	observed (*Fro*)	Z-Statistic	*p-value*
1	30.10%	25.82%	-1.29	0.198
2	17.61%	22.07%	1.62	0.106
3	12.49%	18.31%	2.47	0.014
4	9.69%	15.02%	2.52	0.012
5	7.92%	4.69%	-1.62	0.106
6	6.69%	3.76%	-1.58	0.115
7	5.80%	2.35%	-2.01	0.045
8	5.12%	4.23%	-0.44	0.662
9	4.58%	3.76%	-0.41	0.681

Source: Elaborated by the authors based on the survey data.

Although the conformity analysis proposed by Nigrini (2012), performed using the MAD calculation, showed a slightly lower value of "non-conformity" (0.0347 or 3.47%) when compared to the percentage referring to the total number of new cases (4.39%), the observed distribution was also found to be non-conforming about the expected distribution according to the NB Law, i.e., 0.0347 > 0.015, or even 3.47% > 1.50%.

When analyzing the data in a longitudinal perspective about the first digits of daily totals of new deaths occurring as a result of COVID-19 in Brazil, between 03/17/2020 and 09/24/2020, only the total with first digits equal to 4, 5, 6, and 9 presented relative frequency observed without statistically significant differences. This analysis was made using a significance level of 5% (Z statistic with *p-value* > 0.05) concerning the expected relative frequency according to the NB Law, as presents in Table 5.

Table 5 - Analysis of the first digit of daily totals of new deaths

1st digit	Relative Frequency		Proportion Test	
	expected (*Fre*)	observed (*Fro*)	Z-Statistic	*p-value*
1	30.10%	40.63%	3.10	0.002
2	17.61%	5.73%	-4.23	0.000
3	12.49%	4.17%	-3.38	0.001
4	9.69%	9.38%	-0.03	0.980
5	7.92%	5.21%	-1.26	0.208
6	6.69%	8.33%	0.77	0.443
7	5.80%	11.46%	3.20	0.001
8	5.12%	9.38%	2.51	0.012
9	4.58%	5.73%	0.59	0.556

Source: Elaborated by the authors based on the survey data.

Regarding the overall conformity of the observed relative frequency distribution compared to the expected frequency distribution according to the NB Law, the longitudinal analysis of the first digits referring to the daily totals of new deaths occurring in Brazil as a consequence of COVID-19 showed a MAD of 0.0516 or 5.16%. Therefore, far above the maximum value indicated as critical according to the conformity parameters proposed by Nigrini (2012), that is, 5.16% > 1.5%.

From the application of the Z test, the analysis of the first digit that presented the best individual results per digit was precisely that referring to the series of data corresponding to the daily totals of cumulative deaths by COVID-19 in Brazil. That is, 7 of the first nine digits analyzed (3, 4, 5, 6, 7, 8, and 9) presented observed relative frequency without statistically significant differences at a 5% significance level (Z statistic with *p-value* > 0.05) concerning the expected relative frequency according to the NB Law, as can be seen in Table 6.

Thus, only the observed relative frequencies for the first digits 1 and 2 showed non-significant Z-statistics, for a significance level of 5% (*p-value* < 0.05). However, it should be remembered that, in general terms, the highest expected relative frequency probabilities according to NB Law are those concerning digits 1 and 2, i.e., 30.10% and 17.61%, respectively, which should produce relevant effects on the MAD and conformity levels of this data series.

Table 6 - Analysis of the first digit of daily cumulative death totals

1st digit	Relative Frequency		Proportion Test	
	expected (*Fre*)	observed (*Fro*)	Z-Statistic	*p-value*
1	30.10%	38.54%	2.47	0.013
2	17.61%	10.42%	-2.52	0.012
3	12.49%	8.33%	-1.63	0.102
4	9.69%	7.81%	-0.76	0.449
5	7.92%	8.33%	0.08	0.937
6	6.69%	6.25%	-0.10	0.921
7	5.80%	7.81%	1.04	0.299
8	5.12%	5.73%	0.22	0.826
9	4.58%	6.77%	1.28	0.201

Source: Elaborated by the authors based on the survey data.

Of the four-time series analyzed (longitudinally), the MAD referring to the first digit of the daily totals of deaths accumulated by COVID-19 was the one that showed the lowest levels of non-conformity ("inconformity"), according to the parameters proposed by Nigrini (2012), that is, 0.0303 or 3.03%.

It should be noted that according to NB Law, the occurrence of digits in datasets conforming to those probabilities of distributions described earlier in Table 1 of this study applies to naturally generated data (DRUICĂ *et al.* 2018). However, although the data series analyzed in these studies originate from a recent and natural event, how these data were surveyed and reported comply with the most adverse conditions possible in the COVID-19 pandemic in general. And especially even more diversely in the case of Brazil, where the absence of central coordination caused states and municipalities to adopt distinct coronavirus countermeasures.

Druică *et al.* (2018) warn about the need to include the effect of time and sample size variables on the results of NB Law conformity testing since most studies have.

In this sense, in the second stage of the analyses proposed for this scientific investigation, the MAD and the conformity analysis by Nigrini (2012) were used as parameters to evaluate the degree of conformity between the relative frequencies (%) observed for the analyzed data (Fro) and the expected relative frequency (%) according to the NB Law (Fre).

As already mentioned, the analysis in question was performed transversally and at the state level, analyzing the first digits of the totals reported daily for each Brazilian state and the federal district, between 02/25/2020 and 09/24/2020 for new and accumulated cases of COVID-19. Also, it was analyzed the first digits of the totals reported daily for each Brazilian state and the federal district, between 03/17/2020 (first death in the country) and 09/24/2020, for new and accumulated deaths due to the pandemic of COVID-19 in Brazil.

Due to the volume of data analyzed (27 federal units and their respective daily totals of new cases and accumulated cases of COVID-19 over 213 days, and also those 27 federal units and their daily total of new deaths and accumulated deaths over 192 days), the results of the analysis of the respective MAD were summarized with the help of the statistical-descriptive parameters presented in Table 7.

Table 7 - Analysis of the MAD in relation to the first digits of the daily total for pandemic COVID-19 at the state level

Parameters	New Cases	Accumulated Cases	New Deaths	Accumulated Deaths
Minimum	0.0184	0.0206	0.0170	0.0242
Maximum	0.1553	0.1945	0.2093	0.1945
Amplitude	0.1369	0.1739	0.1923	0.1703
Average	0.0548	0.0565	0.0561	0.0546
Median	0.0486	0.0504	0.0479	0.0508
Standard Deviation	0.0244	0.0294	0.0310	0.0259
Coefficient of variation	44%	52%	55%	47%

Source: Elaborated by the authors based on the survey data.

The mean and median MAD calculated for each day (transversely) at the state level since the beginning of the pandemic of COVID-19 in Brazil corroborated with the values observed for the respective MAD identified from the longitudinal analysis performed at the national level. Since in all four categories of data analyzed, the MAD (mean and/or median) were much higher than the maximum critical value admitted for conformity proposed by Nigrini (2012), i.e., form higher 0.0150 or 1.50%.

This same observation regarding the mean and median values of the MAD described in Table 6 also applies to the minimum values observed for those four categories of data analyzed. It is noteworthy that even the minimum values observed for the MAD of new cases (0.0184 or 1.84%), accumulated cases (0.0206 or 2.06%), news deaths (0.0170 or 1.70%), and accumulated deaths (0.0242 or 2.42%) were higher than the maximum critical value allowed (0.0150 or 1.50%) for conformity, as proposed by Nigrini (2012).

Thus, these results showed that none of the information about the COVID-19 pandemic data in Brazil presented conformity according to the NB Law. Whether at the national or state level, whether from the longitudinal perspective or the cross-sectional perspective.

Regarding results of this nature, i.e., totally out of conformity with the NB Law, it should be emphasized that such deviations should not be taken entirely as frauds or errors. Likewise, full conformity should not be seen as an absence of irregularities, as Costa *et al.* (2012) noted.

Additionally, considering the dispersion of data sources, their characteristics, their levels of geographic aggregation, and, probably the biggest problem, the levels of COVID-19 testing practiced in Brazil, it would be admissible to consider what was called by Druică *et al.* (2018) as a "specific conformity level." That is, it could be admitted that, due to their particular characteristics (logistics, data distortions, forms of geographic aggregation, errors, and/or negligence, among others), the data series analyzed in this research would present different and their levels of conformity, "detaching" from the

probability distribution of frequencies according to the NB Law, as a result of a natural process of generation of these data. In this case, those factors specific to each natural process of data generation would imprint a kind of unique signature on them.

5 Conclusions

Considering the relevance of information systems in the fight against public health emergencies such as the pandemic of COVID-19, and also the possibility of the application of the NB Law to contribute to the evaluation of these systems, this research aimed to evaluate the conformity of the information regarding the number of cases of contamination and deaths by COVID-19 in Brazil according to the NB Law, from the moment of the occurrence of the first case of the disease in the country until September 2020.

Furthermore, considering that most of the studies on the application of the NB Law are based on cross-sectional clippings of data (at a given moment in time) and that there is still little research with applications of this nature that considers the arrangement of the set of variables analyzed over time, this research analyzed both the totals of data arranged overtime at the national level, and their daily behavior at the state level.

In both cases, longitudinal-national and cross-state, the data regarding contamination cases and deaths did not show behavior in conformity with the NB Law.

Since that total lack of conformity has been proven in both data perspectives (longitudinal and cross-sectional), and also in both levels of totalization (national and state), it should be noted that there are already studies that suggest the existence of data-specific levels of conformity due to their particularities.

Thus, it is emphasized that the subject in question is far from being exhausted, which leaves room for continuity and deepening of research in this regard. In this sense, it is suggested to continue this research with both perspectives (longitudinal-national and transversal-state). However, using other metrics of conformity, or even studies that consider as conformity metrics the Z test and the MAD allied to the evaluation proposal presented by Nigrini (2012). Nevertheless, at the municipal level (longitudinal and transversal), comparing with results observed in the present investigation.

Acknowledgments

We thank the reviewers and editors for their comments and suggestions. The last author is grateful for the financial support of CAPES.

CARMO, C. R. S., CANEPPELE, F. L., NUNES, F. C. Análise dos casos de contaminações e mortes por covid-19 no Brasil segundo a lei de Newcomb-Benford. *Rev. Bras. Biom*. Lavras, v.39, n.4, p.522-535, 2021.

- *RESUMO: A utilização da Lei de Newcomb-Benford na avaliação da qualidade de sistemas de informações sanitárias e/ou epidemiológicas pode permitir que se tomem decisões relevantes para a melhoria desses sistemas. Nesse contexto, esta pesquisa teve por objetivo realizar uma avaliação de conformidade das informações referentes às quantidades de casos de contaminação e de mortes por COVID-19 no Brasil segundo a Lei de Newcomb-Benford, desde o momento da ocorrência do primeiro caso da doença e da primeira morte por COVID-19 no país até o mês de setembro de 2020. Com o auxílio de estatísticas descritivas e a utilização de métricas relacionadas ao teste Z e ao desvio absoluto médio foi possível observar que, tanto sob a perspectiva longitudinal-nacional e quanto a perspectiva transversal-estadual, os dados quantitativos referentes aos casos de contaminação pelo coronavírus e aos óbitos acontecidos em decorrência da COVID-19 não apresentaram o comportamento esperado segundo a Lei de Newcomb-Benford. Devido à ausência de conformidade em relação à Lei de Newcomb-Benford, suspeita-se da ocorrência de algum nível de conformidade específico desse tipo de dado, no contexto brasileiro, uma vez que, já existem estudos que sugerem a existência de níveis de conformidade próprios para certos tipos de dados.*
- *PALAVRAS-CHAVE: Epidemiologia; conformidade; métodos quantitativos aplicados.*

References

BARIFOUSE, R. Coronavírus: onde acompanhar os números da pandemia no Brasil após apagão de dados do governo. *BBC News Brasil*, página eletrônica, 8 junho 2020. Available at: https://www.bbc.com/portuguese/brasil-52974181. Accessed on: Sep 24, 2020-14:22.

BENFORD, F. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, v.78, n.4, p.551-572, 1938.

BRASIL, Ministry of Health. *Painel de casos de doença pelo coronavírus 2019 (COVID-19) no Brasil pelo Ministério da Saúde*. Brasília: Ministério da Saúde/Secretaria de Vigilância em Saúde (SVS): Guia de Vigilância Epidemiológica do COVID-19, 2020. Available at: https://covid.saude.gov.br/. Accessed on: Sep 24, 2020-18:31.

BRASIL.IO.COVID-19: *Boletins informativos e casos do coronavírus por município por dia*. 24 set. 2020. Available at: https://brasil.io/dataset/covid19/caso_full/. Accessed on: Sep 24, 2020-21:18.

COSTA, J. I. F.; HENRIQUES, D. B. B.; MELO, S. B.; SANTOS, J. Análise de métodos contabilométricos para determinação de conformidade da Lei Newcomb-Benford aplicados à auditoria contábil. *Revista Gestão Pública: Práticas e Desafio***s**, v.3, n.6, p.292-314, 2012.

DRUICĂ, E.; OANCEA, B.; VÂLSAN, C. Benford's law and the limits of digit analysis. *International Journal of Accounting Information Systems*, v.31, p.75-82, 2018.

GÓMEZ-CAMPONOVO, M.; MORENO, J.; IDROVO, Á. J.; PÁEZ, M.; ACHKAR, M. Monitoring the Paraguayan epidemiological dengue surveillance system (2009-2011) using Benford's law. *Biomédica*, v.36, p.583-592, 2016.

IDROVO, A. J.; MANRIQUE-HERNÁNDEZ, E. F. Data quality of Chinese surveillance of COVID-19: objective analysis based on WHO´s situation reports. *Asia Pacific Journal of Public Health*, v.32, n.4, p.165-167, 2020.

LEE, K-B.; HAN, S.; JEONG, Y. COVID-19, flattening the curve, and Benford's law. *Physica A: Statistical Mechanics and its Applications*, v. 559, artigo 125090, 2020.

MANRIQUE-HERNÁNDEZ, E. F.; FERNÁNDEZ-NIÑO, J. A.; IDROVO, A. J. Global performance of epidemiologic surveillance of Zika virus: rapid assessment of an ongoing epidemic. *Public Health*, n.143, p.14-16, 2017.

MANRIQUE-HERNÁNDEZ, E. F.; MORENO-MONTOYA, J.; HURTADO-ORTIZ, A.; PRIETO-ALVARADO, F. E.; IDROVO, A. J. Desempeño del sistema de vigilancia colombiano durante la pandemia de COVID-19: evaluación rápida de los primeros 50 días. *Biomédica*, v.40 (Supl. 2 - Infecciones respiratórias), p.1-8, 2020.

NEWCOMB, S. Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, v.4, n.1, p. 39-40, 1881.

NIGRINI, M. J. *Benford's Law*: *aplications for forensing accounting, auditing, and fraud detection*. New Jersey: Wiley, 2012.

PAN, S. L.; ZHANG, S. From fighting COVID-19 pandemic to tackling sustainable development goals: An opportunity for responsible information systems research. *International Journal of Information Management*, article 102196, Jul. 2020. Available at: https://doi.org/10.1016/j.ijinfomgt.2020.102196. Accessed on: Sep 28, 2020.

Received on 13.10.2020
Approved after revised on 03.05.2021