Adherence of National and Regional U.S. COVID-19 Case Data with Benford's Law

Tejus Pradeep M.D. [1], Siddharth Arun B.S.[2], Advaitaa Ravipati M.S. [3], Bibhav Poudel B.S.[2]

[1] Department of Internal Medicine, Reading Hospital, Reading, PA, USA
[2] School of Medicine, Johns Hopkins University, Baltimore, MD, USA
[3] School of Medicine, University of Miami, Miami, FL, USA

READING HEALTH SYSTEM
Advancing Health. Transforming Lives.

Introduction

- In the United States, there were 5.2 million cases and 168,000 thousand deaths attributed to Coronavirus disease 2019 (COVID-19) as of August 15, 2020[1]
- Benford's Law (Law of First Digits) - an observation of the expected frequency distribution of the first digit in a set of numbers
- Benford's Law has been validated to detect fraud and accuracy of data collection with several real-world numerical data including population metrics, death rates, physical constants, and mathematical numbers
- Benford's law has previously been used to monitor epidemiological surveillance during other pandemics including the influenza A (H1N1) outbreak[2]
- Application of Benford's law to serve as a quality assurance metric for large datasets has been demonstrated[3]

Objectives

- **Aim:** To compare first-digit distributions of U.S. cumulative cases, both overall and stratified by geographical regions and divisions, to assess for adherence to Benford's distribution in order to assess the accuracy of data collection and reporting

Materials and Methods

- Data about COVID-19 cases and deaths stratified by U.S. counties was obtained on August 15th, 2020 using the Johns Hopkins University national COVID-19 tracker[1]
- Data about cumulative cases and deaths were organized by leading digit of case or death number, aggregated without using time information, and organized by frequency distribution of leading digits stratified by U.S. states, divisions, and region groupings
- All official U.S. case data was included in this study without exceptions

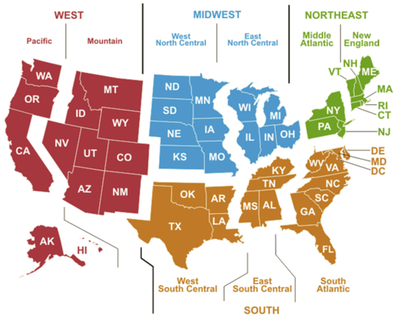


Figure 1. U.S. Census Regional and Divisional Groupings of U.S. States

Materials and Methods cont.

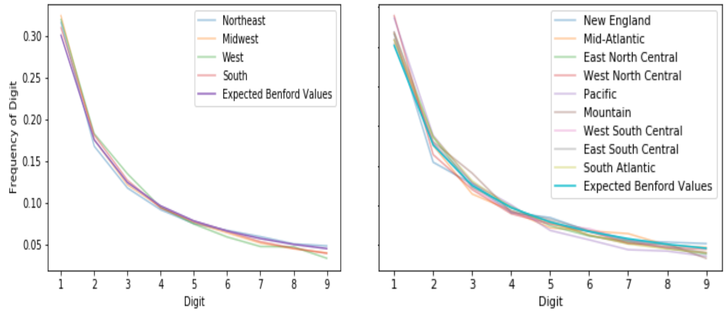
- For large datasets, mean absolute deviation (MAD) and sum of squares difference (SSD) have been validated to be most accurate in comparing expected and actual frequency distributions

$$MAD = \left(\sum_{i=1}^{K} |AP - EP|\right) \div K \qquad SSD = \sum_{i=1}^{K} (AP - EP)^2 \times 10^4$$

- MAD empirically established conformity criteria: close conformity (0.000-0.006), acceptable conformity (0.006-0.012), marginal conformity (0.012-0.015), and non-conformity (values above 0.015)
- SSD empirically established conformity criteria: perfect conformity (0-2), acceptable conformity (2-25), marginal conformity (25-100), non-conformity (values greater than 100).

Results

- The South has had the greatest number of COVID-19 cases (2,405,333) while the Midwest had the least (816,290)
- Overall, regional, and divisional sub-grouped first-digit frequencies of U.S. COVID-19 cases showed significant adherence to Benford's Law.
- Eight geographical divisions showed case numbers with either acceptable or close conformity to Benford's distribution by both MAD and SSD metrics (MAD < 0.012, SSD 25).



Leading Digit	Benford dist.	US Overall	Northeast	Midwest	South	West
# of Cases (Aug 15, 2020)		5,286,519	945,407	816,290	2,405,333	1,119,489
Digit Frequencies 1	0.3010	0.3160	0.3160	0.3242	0.3100	0.3197
2	0.1760	0.1785	0.1684	0.1761	0.1816	0.1833
3	0.1249	0.1263	0.1184	0.1222	0.1276	0.1354
4	0.0969	0.0951	0.0924	0.0958	0.0943	0.0952
5	0.0791	0.0776	0.0759	0.0772	0.0785	0.0753
6	0.0669	0.0654	0.0677	0.0648	0.0666	0.0598
7	0.0579	0.0535	0.0602	0.0526	0.0542	0.0480
8	0.0511	0.0469	0.0515	0.0458	0.0463	0.0484
9	0.0457	0.0404	0.0492	0.0407	0.0403	0.0344
MAD		0.0042 Close conformity	0.0048 Close conformity	0.0052 Close conformity	0.0039 Close conformity	0.0081 Acceptable conformity
SSD		3.09 Acceptable Conformity	3.73 Acceptable Conformity	6.40 Acceptable Conformity	1.95 Perfect Conformity	8.13 Acceptable Conformity

Table 1. Comparison of United States COVID-19 Cases from January 22 - August 15, 2020 to Benford's Law.

Conclusion

1. Data collection and handling of data about COVID-19 case numbers in the United States has been performed appropriately without evidence of manipulation or fraud.
2. We show a novel finding of adherence to Benford's law for COVID-19 cases among U.S. regions and divisions
    a) This finding is especially pertinent given significant regional differences in confirmed COVID-19 cases initially[4], concerns about political partisanship affecting disease surveillance[5], and changes by the Department of Health and Human Services (HHS) regarding rules for hospitals reporting case data to the agency directly rather than to both the CDC and HHS.[6]
    b) We show that per Benford's law analysis, there is no evidence for fraud of manipulation, regionally or nationwide

Limitations

- In this study, we analyzed reported cases, not reported deaths in the United States. Though variations from Benford's Law could occur with death rates, it is harder to differentiate intentional manipulation of data vs. legitimate regional and individual differences in COVID-19 deaths. Further studies are needed in this realm
- Since this is analysis of first-digit frequencies, we cannot comment on future projections or model rates in COVID-19 cases and deaths.
- Since we utilize county level data, we cannot detect individual instances of data manipulation with sufficient power – only general trends in the data

References

1. Johns Hopkins University. COVID-19 United States Cases by County. Accessed: August 15, 2020. https://coronavirus.jhu.edu/us-map
2. Idrovo AJ, Fernández-Niño JA, Bojórquez-Chapela I, Moreno-Montoya J. Performance of public health surveillance systems during the influenza A(H1N1) pandemic in the Americas: testing a new method based on Benford's Law. *Epidemiol Infect*. 2011;139(12):1827-1834. doi:10.1017/S095026881100015X
3. Daniels J, Caetano SJ, Huyer D, et al. Benford's Law for Quality Assurance of Manner of Death Counts in Small and Large Databases. *J Forensic Sci*. 2017;62(5):1326-1331. doi:10.1111/1556-4029.13437
4. Geographical Differences in COVID-19 Cases, Deaths, and Incidence. Morbidity and Mortality Weekly Report (MMWR). Centers for Disease Control and Prevention. April 2020; 69(15);465-471. Accessed August 24, 2020.
5. Tyson A. Republicans Remain far less likely than Democrats to view COVID-19 as a major threat to public health. Pew Research Center. July 22, 2020. Accessed August 24, 2020.
6. Florko N. How HHS's new hospital data reporting system will actually affect the U.S. COVID-19 response. STAT News. July 16, 2020. Accessed August 24, 2020.