ZIBELINE INTERNATIONAL™ PUBLISHING

Matrix Science Mathematic (MSMK)

DOI: http://doi.org/10.26480/msmk.02.2021.35.38

ISSN: 2521-0831 (Print)
ISSN: 2521-084X (Online)
CODEN: MSMADH

RESEARCH ARTICLE

METRICS IN SMALL-SIZED QURAN DATASET FOR BENFORD'S LAW

M. Z. A. M. Jaffar*, A. N. Zailan, N. H. Izamuddin

Department of Mathematics and Statistics, Faculty of Science, Universiti Putra Malaysia 43400 Serdang, Selangor, Malaysia.
****Corresponding Author email: maizurwatul@upm.edu.my***

ARTICLE DETAILS

Article History:

Received 25 September 2021
Accepted 03 November 2021
Available online 22 November 2021

ABSTRACT

Benford's law is widely applied in testing anomalies in various dataset, including accounting fraud detection and population numbers. It is a statistical regularity, which is said that it works better with larger datasets that span large orders of magnitude distributed in a non-uniform way. In this study, we examine the potential metrics in small-sized Quran dataset that are applicable for the Benford's law. Against our expectations, we find that the Quran dataset conforms to the Benford's law. We provide evidence that metrics such as total paragraph per chapter and total verse per chapter conform to Benford's distribution. However, total verse is closer to Benford's law prediction compared to total paragraph.

KEYWORDS

Benford's law; Quran; Small-sized dataset; Metric.

1. INTRODUCTION

The Quran is an Islam religious text that includes God's message delivered to the Prophet Muhammad S.A.W by Gabriel, an angel, to be recited, comprehended, and practised as a guidance or living style for humanity (Oktaviani et al., 2019). The sacred Quran comprises 114 surahs and approximately 6,236 verses incorporating 77,477 terms, and each verse and term is addressed by interpretation, parse, and explanation (Hegazi et al., 2015). Good clarification of the Quran involves getting accustomed to passages in the Quran. Al-Khatib al-Iskafi asserted that only 28 or roughly 25% of the 114 Quran's chapters do not include identical or repetitive passages (Oktaviani et al., 2019). Benford's Law (BL), sometimes identified as the first digit law, describes how integers are distributed in massive databases. This rule considers that the regularity of the first integers of numbers is not uniformly scattered across lots of naturally occurring structures. Zipf's law is a similar empirical rule. BL, in reality, may be viewed as a particular instance of Zipf's law. Zipf verified that, provided a database containing a language's frequent term, the occurrence of each term is inversely related towards its place in the ordering of term's regularity (Melián et al., 2017).

2. LITERATURE REVIEW

2.1 Benford's Law

2.1.1 Pioneer

The renowned Benford's Law (BL) was first postulated by Astronomer Simon Newcomb in 1881 and was established by Frank Benford in 1938 (Melita and Miraglia, 2021). Newcomb and subsequently Benford discovered that the incidence of significant digits across colossal data is not homogeneous but instead follows a logarithmic trend, with lower integers appearing first more often than bigger digits (Ausloos et al., 2014). In 1881, Simon Newcomb noticed that logarithmic tables favour smaller digits in the leading spot. In his publication "Note on the Frequency of Use of the Different Digits in Natural Numbers", he detailed his findings (Druică et al., 2018). Frank Benford, a physicist, published "The Law of Anomalous Numbers" in 1938, wherein he discovered the worn-out pages of logarithmic tables, similar to Newcomb. Benford expanded his study, unknown of Newcomb's work, by collecting more than 20,000 data from various resources to evaluate the probability of occurrence for every leading digit. Benford's Law was born from the findings of this research (Chang, 2017).

2.2 Definition and Concepts

Benford's Law is a rule of phenomenology that describes the probability distribution of a data collection's first significant digits (Shi et al., 2017). BL states that lower-order values such as 1, 2, and 3 are more common than higher-order values (Máté et al., 2017). Newcomb-Benford's Law, recognised as "the first significant digit law" or "the law of anomalous numbers," is rooted in a finding in which a rule concerning the circulation of the leading positive significant numbers in numeric values was formalised, where the likelihood of the most significant digits are dispersed unevenly (Palacios, 2020). It emphasises that in a vast data collection, the occurrence of dispersion of the first significant integer meets the respective equation:

$$P_{\eta_1}^{B} = \log_{10}(1+\eta_1) - \log_{10}(\eta_1) \quad (1)$$

Quick Response Code	Access this article online	
	Website: www.matrixsmathematic.com	**DOI:** 10.26480/msmk.02.2021.35.38

Cite the Article: M. Z. A. M. Jaffar, A. N. Zailan, N. H. Izamuddin (2021). Metrics in Small-Sized Quran Dataset for Benford's Law. *Matrix Science Mathematic, 5(2): 35-38.*

where η_1 is a provided digit, η's first non-zero number while $P_{\eta_1}^{B}$ is the Benford's likelihood of that value to occur (Melita and Miraglia, 2021).

Benford showed further that "first digit law" applies to almost every set of values in a given dataset. Random datasets that have constraints, such as winning digits, contact information, or fuel costs, are examples of exclusions (Chang, 2017). The fundamental features of Benford's Law are scaling and basis invariance, with the scale-invariant characteristic implying that BL remains valid although the measurement's units are altered. That is to say, the degree to which certain data fits Benford's Law is unaffected by the measuring structure (Badal-Valero et al., 2018).

2.3 Application

Awareness of Benford's Law has increased throughout a variety of uses, such as detecting fraud, system programming, and information extraction, with the evolution of digital technologies and the capacity to analyse enormous data sets (Chang, 2017). It was used to compare the first and second value possibilities to determine the weights, orbiting durations, semimajor axial, eccentricities, and radius of current exoplanets (Melita and Miraglia, 2021). In addition, BL testing was used to see if the worldwide market price of certain financial data gathered by the Financial Times Security Exchange had any inaccurate readings (Shi et al., 2017). One of the disciplines that profited enormously from such discoveries was fraud detection, where BL began to be utilised as the foundation of audits (Druică et al., 2018). In order to detect academic fraud, co-authors and publishers should start introducing an analysis program that uses Benford Law to identify possible "warning sign" publications in order to reduce the likelihood of fraud and thereby improve the reputation of academic research papers (Horton et al., 2020). In the framework of an actual Spanish legal case, we use BL and artificial learning algorithms to identify trends of tax evasion offenders (Badal-Valero et al., 2018). On the other hand, adherence of social networking algorithms and Intelligence Analysis acts to Benford's Law was inspected, where the findings revealed that bots obey BL, implying that utilising this rule can aid in the detection of harmful online programmed entities and associated behaviours on media platforms (Madahali and Hall, 2020).

2.4 Claims of Benford's law applied on large datasets

Benford's Law could not operate given tiny datasets, and a sampling of 200 observations for every order or more of magnitude is necessary (Melita and Miraglia, 2021). Furthermore, it may be utilised as a technique for identifying abnormalities in massive datasets, which can be implemented to subgroups of more extensive sets to limit the range of probable unusual information and make the approach easier to manage. The implications of tiny datasets for BL are almost entirely excluded. (Druică et al., 2018). All numbers from 1 to 9 might have an equal chance of appearing as the first number in any practical randomised assessment if the numbers are scattered evenly. That's not the reality since this rule defies logic and seems to be applicable to enormous amounts of data with ease (D'Alessandro, 2020).

3. METHODOLOGY

The Quran contains 114 chapters (surahs), where each chapter is divided into verses (ayats). The chapters are not equal in length. For example, Surah Al-Kawthar has only three verses, which is the shortest chapter, while the longest chapter is Al-Baqarah contains 286 verses.

The Quran dataset identified in our study based on total Chapter categorised by:

- Total verse per Chapter
- Total paragraph per Chapter

This study applied the Benford's law for the two Quran datasets mentioned above in Microsoft Excel, where results are shown in Tables 1 and 2 and Charts 1 and 2.

Table 1: Frequency percentage of first digit of total verse in the Quran is compared with frequency percentage of first digit predicted by the Benford's law.

First Digit	*First Digit Frequency of Total Verse*	*% First Digit Frequency of Total Verse*	*% Benford's law Prediction*
1	30	26.32%	30.10%
2	17	14.91%	17.61%
3	12	10.53%	12.49%
4	11	9.65%	9.69%
5	14	12.28%	7.92%
6	7	6.14%	6.69%
7	8	7.02%	5.80%
8	10	8.77%	5.12%
9	5	4.39%	4.58%
Sum	***114***	***100.00%***	

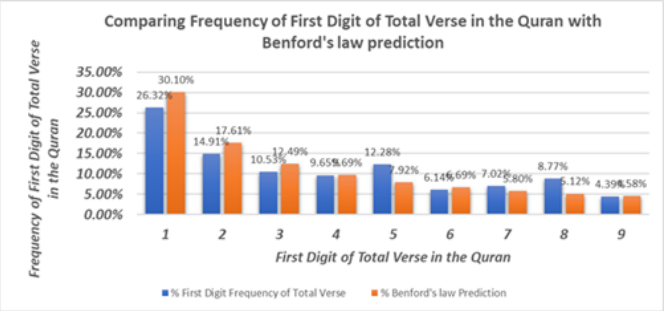

Figure 1: This chart shows frequency percentage of first digit of total verse in the Quran and frequency percentage of first digit predicted by the Benford's law.

Next table and chart are result related to total paragraph dataset.

Table 2: Frequency percentage of first digit of total paragraph in the Quran is compared with frequency percentage of first digit predicted by the Benford's law.

First Digit	*Frequency of First Digit of Total Paragraph*	*% Frequency of First Digit of Total Paragraph*	*% Benford's law Prediction*
1	47	41.23%	30.10%
2	26	22.81%	17.61%
3	10	8.77%	12.49%
4	7	6.14%	9.69%
5	5	4.39%	7.92%
6	8	7.02%	6.69%
7	5	4.39%	5.80%
8	2	1.75%	5.12%
9	4	3.51%	4.58%
Sum	***114***	***100.00%***	

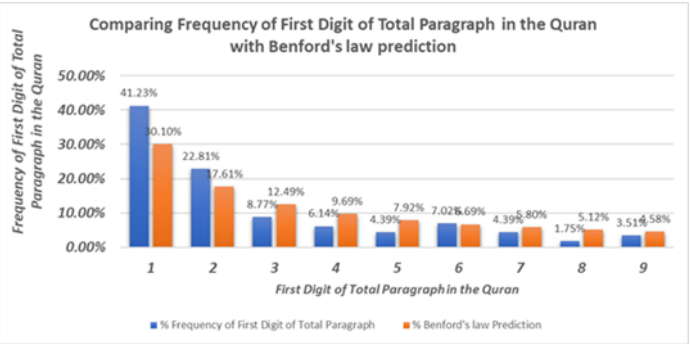

Figure 2: This chart shows frequency percentage of first digit of total paragraph in the Quran and frequency percentage of first digit predicted by the Benford's law.

4. Discussion

From Tables 1 and 2 and Figure 1 and 2, dataset of total verse is closer to Benford's law prediction compared to total paragraph. Subsequently, we showed such a closeness mathematically by calculating sums of the squares of the differences between frequency precentage of the datasets and frequency percentage predicted by the Benford's law.

Based on Table 1, we calculated sum of the squares of the differences between frequency of total verse and the Benford's law prediction as shown in Table 3.

Table 3: Sum of the squares of the differences between frequency of total verse and the Benford's law prediction.

% First Digit Frequency of Total Verse	% Benford's law Prediction	Sum of the squares of the differences
26.32%	30.10%	0.14%
14.91%	17.61%	0.07%
10.53%	12.49%	0.04%
9.65%	9.69%	0.00%
12.28%	7.92%	0.19%
6.14%	6.69%	0.00%
7.02%	5.80%	0.01%
8.77%	5.12%	0.13%
4.39%	4.58%	0.00%

Based on Table 2, we calculated sum of the squares of the differences between frequency of total paragraph and the Benford's law prediction as shown in Table 4.

Table 4: Sum of the squares of the differences between frequency of total paragraph and the Benford's law prediction.

% Frequency of First Digit of Total Paragraph	% Benford's law Prediction	Sum of the squares of the differences
41.23%	30.10%	1.24%
22.81%	17.61%	0.27%
8.77%	12.49%	0.14%
6.14%	9.69%	0.13%
4.39%	7.92%	0.12%
7.02%	6.69%	0.00%
4.39%	5.80%	0.02%
1.75%	5.12%	0.11%
3.51%	4.58%	0.01%

5. Conclusion

This study shows that total verse and total paragraph in the Quran are suitable metrics for the Benford's law. Furthermore, comparing between the metrics, total verse presents a better fit to the Benford's law as compared to total paragraph. This is because its sum of the squares of the differences is closer to zero, which means closer gap between the two values.

Acknowledgement

We would like to thank Universiti Putra Malaysia.

References

Adel, M.E., 2021. Zipf's law applications in patent landscape analysis. *World Patent Information,* 64, Pp. 102012. https://doi.org/10.1016/j.wpi.2020.102012

Aitchison, L., Corradi, N., Latham, P.E., 2016. Zipf's Law Arises Naturally When There Are Underlying, Unobserved Variables. *PLoS Computational Biology,* 12 (12), Pp. e1005110. https://doi.org/10.1371/journal.pcbi.1005110

Arshad, S., Hu, S., Ashraf, B.N., 2018. Zipf's law, the coherence of the urban system and city size distribution: Evidence from Pakistan. *Physica A: Statistical Mechanics and its Applications,* 513, Pp. 87-103. https://doi.org/10.1016/j.physa.2018.08.065

Ausloos, M., Herteliu, C., Ileanu, B., 2014. Breakdown of Benford's Law for Birth Data. *Physica A: Statistical Mechanics and its Applications,* 419 (1). http://dx.doi.org/10.1016/j.physa.2014.10.041

Avetisyan, S., 2019. Why is Zipf's law important for cities? https://doi.org/10.13140/RG.2.2.34651.21284

Badal-Valero, E., José, A., Alvarez-Jareño, Jose, M.P., 2018. Combining Benford's Law and machine learning to detect money laundering. An actual Spanish court case. *Forensic Science International,* 282, Pp. 24-34. https://doi.org/10.1016/j.forsciint.2017.11.008

Chang, J.C., 2017. A Study Of Benford's Law, With Applications To The Analysis Of Corporate Financial Statements [Master's Thesis]. ETDA. https://etda.libraries.psu.edu/files/final_submissions/14504

Chen, Y., 2020. Exploring the level of urbanization based on Zipf's scaling exponent. *Physica A: Statistical Mechanics and its Applications,* 566, Pp. 125620. https://doi.org/10.1016/j.physa.2020.125620

Cong, M., Li, C., Ma, B., 2019. First digit law from Laplace transform. *Physics Letters A,* 383, Pp. 1836-1844. https://doi.org/10.1016/j.physleta.2019.03.017

Cristelli, M., Batty, M., Pietronero, L., 2012. There is More than a Power Law in Zipf. *Scientific Reports, 2* (812). https://doi.org/10.1038/srep00812

D'Alessandro, A., 2020. Benford's law and metabolomics: A tale of numbers and blood. *Transfusion and Apheresis Science, 59* (2020), Pp. 103019. https://doi.org/10.1016/j.transci.2020.103019

Druică, E., Oancea, B., Vâlsan, C., 2018. Benford's law and the limits of digit analysis. *International Journal of Accounting Information Systems,* 31 (8), Pp. 75-82. https://doi.org/10.1016/j.accinf.2018.09.004

Ferrer-i-Cancho, R., Vitevitch, M.S., 2018. The origins of Zipf's meaning-frequency law. *Journal of the American Society for Information Science and Technology,* 69 (11), Pp. 1369-1379. https://doi.org/10.1002/asi.24057

Glattfelder, J.B., 2018. *Information–Consciousness–Reality. How a New Understanding of the Universe Can Help Answer Age-Old Questions of Existence.* Springer Open. https://doi.org/10.1007/978-3-030-03633-1

Hegazi, M.O., Hilal, A., Alhawarat, M., 2015. Fine-Grained Quran Dataset. *International Journal of Advanced Computer Science and Applications (IJACSA),* 6 (12), Pp. 308-313. https://dx.doi.org/10.14569/IJACSA.2015.061241

Horton, J., Kumar, D.K., Wood, A., 2020. Detecting academic fraud using Benford law: The case of Professor James Hunton. *Research Policy,* 49, Pp. 104084. https://doi.org/10.1016/j.respol.2020.104084

Lestrade, S., 2017. Unzipping Zipf's law. *PLoS ONE,* 12 (8), Pp. e0181987. https://doi.org/10.1371/journal.pone.0181987

Li, W., 2002. Zipf's Law Everywhere. *Glottometrics, 5*, Pp. 14-21. https://www.researchgate.net/publication/253290454

Madahali, L., Hall, M., 2020. Application of the Benford's law to Social bots and Information Operations activities. *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)* (pp. 1-8). IEEE Conference Proceedings (IEEE Conf Proc). http://dx.doi.org/10.1109/CyberSA49311.2020.9139709

Máté, D., Sadaf, R., Tarnóczi, T., Fenyves, V., 2017. Fraud Detection by Testing the Conformity to Benford's Law In The Case Of Wholesale Enterprises. *Polish Journal of Management Studies, 16* (1), Pp. 115-126. https://doi.org/10.17512/pjms.2017.16.1.10

Melián, J.A.P., Conejero, J.A., Ferri, C., 2017. Zipf's and Benford's laws in Twitter hashtags. *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 84-93. https://aclanthology.org/E17-4009

Melita, M.D., Miraglia, J.E., 2021. On the applicability of Benford law to exoplanetary and asteroid data. *New Astronomy,* 89, Pp. 101654. https://doi.org/10.1016/j.newast.2021.101654

Oktaviani, D., Bijaksana, M.A., Asror, I., 2019. Building a Database of Recurring Text in the Quran and its Translation. *Procedia Computer Science,* 157, Pp. 125-133. https://doi.org/10.1016/j.procs.2019.08.149

Palacios, N.T., 2020. Benford's Law. History, mathematical justification and applications [Final Degree Project]. Universidad de Valladolid. http://uvadoc.uva.es/handle/10324/43776

Powers, D.M.W., 1998. Applications and Explanations of Zipf's Law. *New Methods in Language Processing and Computational Natural Language Learning.* https://aclanthology.org/W98-1218/

Shi, J., Ausloos, M., Zhu, T., 2017. Benford's law first significant digit and distribution distances for testing the reliability of financial reports in developing countries. *Physica A: Statistical Mechanics and its Applications,* 492, Pp. 878-888. https://doi.org/10.1016/j.physa.2017.11.017

Shyklo, A.E., 2017. Simple explanation of Zipf's mystery via new rank-share distribution, derived from combinatorics of the ranking process. *Combinatorics of the Ranking Process*. http://dx.doi.org/10.2139/ssrn.2918642.