Stylometry and Numerals Usage: Benford's Law and Beyond

Subjects: Language & Linguistics
Contributor: Andrei Zenkov

Benford's Law is a strange manifestation of the law of large numbers (understood as the combined action of a large number of random factors leading to a result that is almost independent of the case).

Keywords: Benford's Law ; first significant digit ; numerals in texts

1. Introduction

Benford's Law is sometimes rightfully called curious, surprising and mysterious [1][2]. There is still no complete understanding of why some data sets obey this law, while others do not. The famous Hill's theorem [3], which gives *sufficient* conditions for the appearance of Benford's Law, does not give the *necessary* ones nor the insight into why and when Benford's Law applies.

Incomplete understanding does not prevent the emergence of more and more proposals for the practical use of Benford's Law in a wide area of sciences from geodesy [4] and geology [5] through genomics [6] and ecology [7][8] to scientometrics [9].

The primary goal of these attempts is to detect various *falsifications* (in a broad sense) and anomalies in data sets [10][11][12]. The questions addressed extends from the possibility of finding the Dyson spheres (presumably built by advanced civilizations) through anomalies in the star emission spectra to the prosaic falsifications of election results [13][14] and financial statements. In the USA, evidence based on Benford's Law [15] has been admitted in criminal cases of financial fraud at the federal, state, and local levels. The Internal Revenue Service of the US federal government, which is responsible for collecting taxes, has been using it for decades to ferret out fraudsters.

2. Benford's Law and Texts

There have been few attempts to link Benford's Law with text data analysis. The first belongs to Benford himself. In his classic work [16], he analyzed

- Arabic numbers (not spelled out) of consecutive front-page news items of a newspaper. "Dates were barred as not being variable, and the omission of spelled-out numbers restricted the counted digits to numbers 10 and over";
- The first 342 street addresses given in an *American Men of Science* edition;
- Numeral usage (except for dates and page numbers) of an issue of the *Readers' Digest*.

Benford stated that fully random data (the first and second items) had an excellent agreement with the "logarithmic law" (It may be explained by Hill's theorem [3]), and the third item was also in agreement with it.

An excellent introduction to Benfordology in general and to the analysis of the use of numerals in *texts* is contained in the paper by Hungerbühler [17].

2.1. Distribution of the First Significant Digits of the Numerals in Compiled Texts

The conditions of Hill's theorem [3] are best fulfilled for *compiled* texts consisting of fragments by *different* authors. In this case, the authors' peculiarities of the texts are averaged, and the distribution is obtained that resembles Benford's one, but with a faster decrease in frequency; the occurrence of digit 1 is significantly higher than expected according to Benford's Law. Starting with digit 3, the actual frequency usually decreases faster than the theoretical one.

In **Figure 1**, the results of the analysis of the compiled text of three collections of Russian-language literary prose are presented:

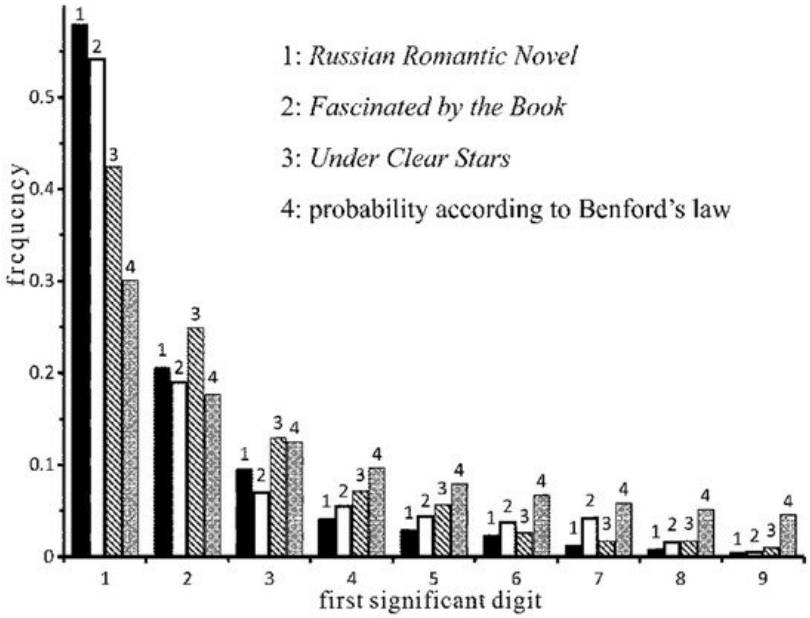

Figure 1. The frequency distribution of the first significant digits of numerals in three collections of Russian-language literary texts. Results are compared with those prescribed by Benford's Law.

- "Russian Romantic Novel" [18];
- "Fascinated by the Book" [19];
- "Under Clear Stars" [20].

For each compilation, the frequency gradually decreases; patterns for different compilations are generally similar, the differences may be related to the peculiarities of the texts in each collection (for example, genre and time of creation, but this requires additional research).

In **Figure 2**, similar results for English-language compilations [21][22][23][24][25][26][27][28] are presented.

Figure 2. The distribution of the first significant digits of numerals in eight collections of English-language literary texts.

2.2. Coherent Literary Texts: The Author's Peculiarities

As a rule, in texts written by *one* author, stable peculiarities in the statistics of the first significant digits are observed.

The results of the analysis of the most voluminous novels by L. Tolstoy (**Figure 3**), F. Dostoevsky (works Nos. 1–9 in **Figure 4**), and I. Goncharov (**Figure 5**) are presented.

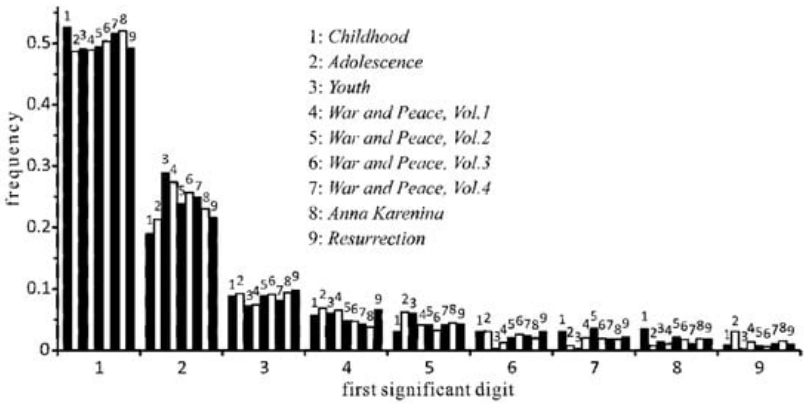

Figure 3. The distribution of the first significant digits of numerals in the works by L. Tolstoy.

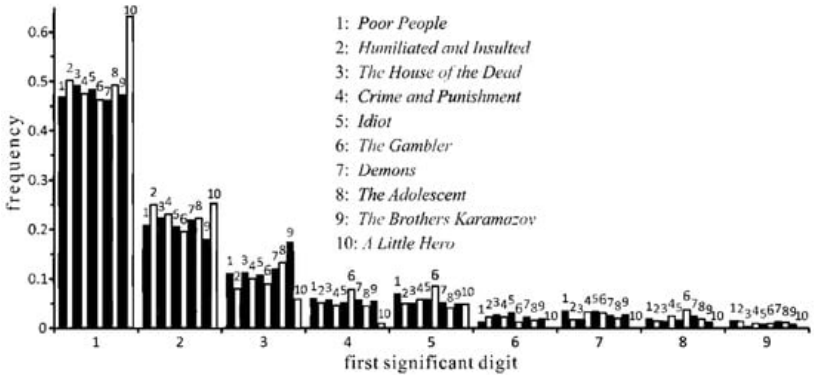

Figure 4. The distribution of the first significant digits of numerals in Dostoevsky's texts. In addition to voluminous works (Nos. 1–9), a shorter one (No. 10) was analyzed for comparison.

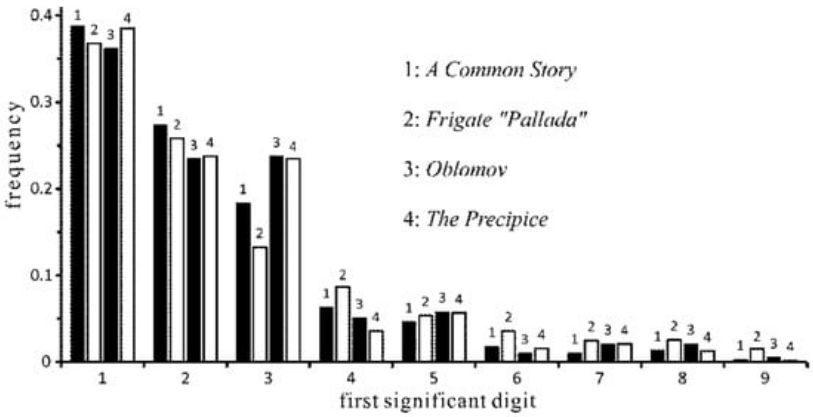

Figure 5. The distribution of the first significant digits of numerals in Goncharov's texts.

2.3. First Significant Digits and Texts Authorship Attribution

The problem of "*And Quiet Flows the Don*"

A well-known problem of the text attribution is the question of the authorship of the novel *And Quiet Flows the Don* and, more broadly, of the entire M. Sholokhov's literary heritage. There are strong arguments for and against plagiarism. The linguistic and statistical study of the novel revealed an extremely heterogeneous text. Many different candidates were put forward for the role of the true authors of its eight parts. There are also doubts about the authorship not only of *And Quiet Flows the Don* but also of the subsequent novels *Virgin Soil Upturned* and *They Fought for Their Country* [29][30].

3. Beyond the Benford's Law

Nigrini [15] has elaborated an express technique of Benford's Law-based accounting fraud detection and an enlarged one.

3.1. The Extension of the Numerals Analysis. Dobychin vs. Platonov

The literary texts of L.I. Dobychin and A.P. Platonov are distinguished by sharp stylistic originality; one finds common literary sources in Russian fiction and analogues in foreign literature [31]. **Figure 6** shows the frequency distribution of the first significant digits of the numerals occurring in the most voluminous works by Dobychin and Platonov [32].

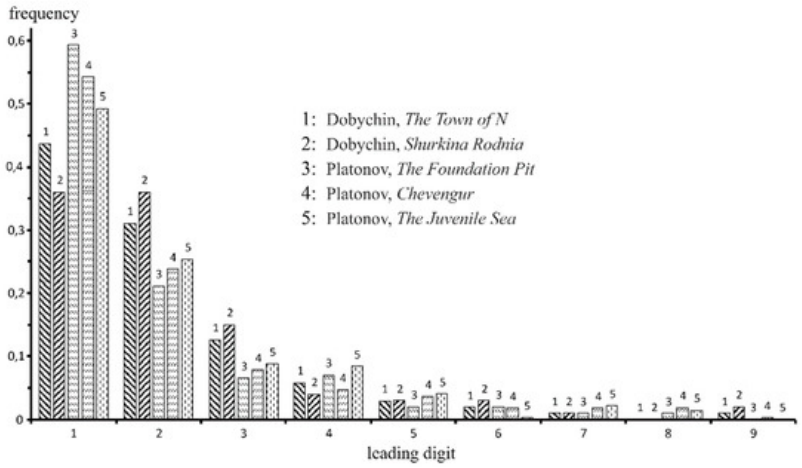

Figure 6. Distribution of relative occurrence frequencies of the first significant digits of numerals in the texts by L. Dobychin and A. Platonov.

3.2. Who Wrote "*The Twelve Chairs*"?

The literary work of popular Soviet authors of the 1920s and 1930s I. Ilf and E. Petrov has repeatedly become the subject of discussion. The novels *The Twelve Chairs* and *The Little Golden Calf* are full of literary allusions; thematically and stylistically, they are related to the texts by V. Kataev, M. Bulgakov, Yu. Olesha, and others [33]. There is nothing comparable to these two works in the literary heritage of Ilf and Petrov. According to the radical point of view [34], Ilf and Petrov are the fake authors of *The Twelve Chairs* and *The Little Golden Calf*, and they were ghosted by Bulgakov.

The numerals extracted from the texts are displayed on frequency graphs [35], which made it possible to directly draw conclusions about the author's style. Information about numerals found in texts was also systematized using the hierarchical cluster analysis [36]. The farthest neighbor clustering was used (which exaggerates differences yet provides clearly defined clusters).

References

1. Fewster, R.M. A Simple Explanation of Benford's Law. Am. Stat. 2009, 63, 29–32.
2. Blondeau Da Silva, S. Limits of Benford's law in experimental field. Int. J. Appl. Math. 2020, 33, 685–695.
3. Hill, T.P. A Statistical Derivation of the Significant-Digit Law. Stat. Sci. 1995, 10, 354–363.
4. Alipour, A.; Alipour, S. Application of Benford's Law in Analyzing Geotechnical Data. Civ. Eng. Infrastruct. J. 2019, 52, 3 23–334.
5. Mangoua, M.J.; Kouassi, K.A.; Douagui, G.A.; Savané, I.; Biémi, J. Application of Benford's Law to Hydrogeological Par ameters: Case of the Baya Watershed (Eastern Côte d'Ivoire). Asian J. Geol. Res. 2019, 2, 1–7.
6. Morag, S.; Salmon-Divon, M. Characterizing Human Cell Types and Tissue Origin Using the Benford Law. Cells 2019, 8, 1004.
7. Özkundakci, D.; Pingram, M. Nature favours "one" as the leading digit in phytoplankton abundance data. Limnologica 2 019, 78, 125707.
8. Cole, M.A.; Maddison, D.J.; Zhang, L. Testing the emission reduction claims of CDM projects using the Benford's Law. Clim. Chang. 2020, 160, 407–426.
9. Vellwock, A.E.; Wei, A. On the Benfordness of Academic Citations. November 2020. Available online: https://www.resea rchgate.net/publication/345437332_On_the_Benfordness_of_academic_citations (accessed on 29 October 2021).
10. Sambridge, M.; Jackson, A. Spotlight on figures for COVID-19. Nature 2020, 581, 384.

11. Farhadi, N. Can we rely on COVID-19 data? An assessment of data from over 200 countries worldwide. Sci. Prog. 202 1, 104, 1–19.
12. Grammatikos, T.; Papanikolaou, N.I. Applying Benford's Law to detect accounting data manipulation in the banking indu stry. J. Financ. Serv. Res. 2021, 59, 115–142.
13. Dacey, J. Benford's Law and the 2020 US Presidential Election: Nothing Out Of The Ordinary. Available online: https://p hysicsworld.com/a/benfords-law-and-the-2020-us-presidential-election-nothing-out-of-the-ordinary/ (accessed on 29 Oc tober 2021).
14. Kossovsky, A.E.; Miller, S.J. Report on Benford's Law Analysis of 2020 Presidential Election Data. Available online: http s://web.williams.edu/Mathematics/sjmiller/public_html/KossoskyMiller_FinalBenfordAnalysis.pdf (accessed on 29 Octob er 2021).
15. Nigrini, M.J. Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection; John Wiley & Sons: H oboken, NJ, USA, 2012; 330p.
16. Benford, F. The law of anomalous numbers. Proc. Am. Philos. Soc. 1938, 78, 551–572.
17. Hungerbühler, N. Benfords Gesetz über führende Ziffern: Wie die Mathematik Steuersündern das Fürchten lehrt. Educ ETH, Publikation der Eidgenössischen Technischen Hochschule Zürich. 2007. Available online: https://ethz.ch/content/ dam/ethz/special-interest/dual/educeth-dam/documents/Unterrichtsmaterialien/mathematik/Benfords%20Gesetz%20üb er%20führende%20Ziffern%20(Artikel)/benford.pdf (accessed on 29 October 2021).
18. Pogorelsky, A.; Titov, V.; Pogodin, M.; Melgunov, N.; Baratynsky, E.; Bestuzhev (Marlinsky), A.; Polevoy, N.; Zagoskin, M.; Rostopchina, E.; Olin, V.; et al. Russian Romantic Novel; Khudozhestvennaia Literatura Publ.: Moscow, Russia, 19 89; 384p. (In Russian)
19. Novikov, N.; Radishchev, A.; Strakhov, N.; Berezaysky, B.; Karamzin, N.; Zhukovsky, V.; Yakovlev, P.; Pushkin, A.; Odoy evsky, V.; Herzen, A.; et al. Fascinated by the Book. Russian Writers on Books, Reading, Bibliophiles; Kniga Publ.: Mos cow, Russia, 1982; 287p. (In Russian)
20. Gorky, A.; Romanov, P.; Tikhonov, N.; Fadeev, A.; Kaverin, V.; Nikulin, L.; Babel, I.; Kolosov, M.; Lavrenev, B.; Sokolov-Mikitov, I.; et al. Under Clear Stars. The Soviet Story of the Thirties; The Moscow Worker Publ.: Moscow, Russia, 1983; 130p. (In Russian)
21. Morris, G.P.; Poe, E.A.; Kirkland, C.M.S.; Leslie, E.; Curtis, G.W.; Hale, E.E.; Holmes, O.W.; Twain, M.; Edwards, H.S.; Johnston, R.M.; et al. The Best American Humorous Short Stories; The Project Gutenberg eBook: Salt Lake City, UT, U SA; Available online: http://www.gutenberg.org/files/10947 (accessed on 10 October 2021).
22. Irving, W.; Poe, E.A.; Hawthorne, N.; Bret Harte, F.; Stevenson, R.L.; Kipling, R. The Short-Story; Transcribed from the 1916 Allyn and Bacon edition; The Project Gutenberg eBook: Salt Lake City, UT, USA; Available online: http://www.gute nberg.org/files/21964 (accessed on 10 October 2021).
23. Kipling, R.; Conan Doyle, A.; Castle, E.; Weyman, S.J.; Collins, W.; Stevenson, R.L. The Lock and Key Library, Classic Mystery and Detective Stories; Transcribed from the 1909 Review of Reviews Co. edition; The Project Gutenberg eBoo k: Salt Lake City, UT, USA; Available online: http://www.gutenberg.org/files/2038 (accessed on 10 October 2021).
24. Johnson, S.; Walpole, H.; Beckford, W. Shorter Novels, Eighteenth Century. The History of Rasselas, The Castle of Otr anto, Vathek; Transcribed from the 1903 Aldine House edition; The Project Gutenberg eBook: Salt Lake City, UT, USA; Available online: http://www.gutenberg.org/files/34766 (accessed on 10 October 2021).
25. Boswell, J.; Wordsworth, W.; Scott, W.; Coleridge, S.T.; Southey, R.; Landor, W.S.; Lamb, C.; Hazlitt, W.; De Quincey, T.; Lord Byron, P.B.; et al. The Best of the World's Classics, Vol. V (of X)—Great Britain and Ireland; Transcribed from t he 1909 Funk & Wagnalls Co. edition; The Project Gutenberg eBook: Salt Lake City, UT, USA; Available online: http://w ww.gutenberg.org/files/22182 (accessed on 10 October 2021).
26. Defoe, D.; Hogg, J.; Irving, W.; Hawthorne, N.; Poe, E.A.; Brown, J.; Dickens, C.; Stockton, F.R.; Twain, M.; Bret Harte, F.; et al. The Great English Short-Story Writers, Volume 1; Transcribed from the 1910 Readers' Library edition; The Proj ect Gutenberg eBook: Salt Lake City, UT, USA; Available online: http://www.gutenberg.org/files/10135 (accessed on 10 October 2021).
27. Dickens, C.; Collins, W.; Gaskell, E.; Procter, A.A. A House to Let; Transcribed from the 1903 Chapman and Hall editio n; The Project Gutenberg eBook: Salt Lake City, UT, USA; Available online: http://www.gutenberg.org/files/2324 (acces sed on 10 October 2021).
28. Blackwood, A.; Rhodes James, M.; Rickford, K.; Harvey, W.F.; Adams Cram, R.; Stevenson, R.L.; Steele, W.D. Masterp ieces of Mystery, Volume 1, Ghost Stories; Transcribed from the 1920 Doubleday, Page & Co. edition; The Project Gute nberg eBook: Salt Lake City, UT, USA; Available online: http://www.gutenberg.org/files/27722 (accessed on 10 October 2021).

29. Kjetsaa, G.; Gustavsson, S.; Beckman, B.; Gil, S. The Authorship of the Quiet Don. Slavica Norvegica; Solum Forlag: O slo, Norway; Humanities Press: Atlantic Highlands, NJ, USA, 1984; Volume 1, 153p.
30. Kuznetsov, F.F. (Ed.) New on Mikhail Sholokhov: Research and Materials; Institute of World Literature: Moscow, Russi a, 2003; 450p. (In Russian)
31. Eidinova, V.V.; Platonov, A.; Dobychin, L. Stylistic Convergence and Repulsion. Andrei Platonov's "Land of Philosopher s": Problems of Creativity. In Proceedings of the International Scientific Conference Dedicated to the 50th Anniversary o f A. Platonov's Death, Moscow, Russia, 23–25 April 2001; pp. 211–219. (In Russian).
32. Zenkov, A.V. Statistics of Numerals in the Text: Development of a New Method of Stylometry, Advances in Economics, Business and Management Research. In Proceedings of the First International Volga Region Conference on Economic s, Humanities and Sports FICEHS 19, Kazan, Russia, 24–25 September 2019; Atlantis Press: Amsterdam, The Netherl ands, 2020; Volume 114, pp. 448–451.
33. Ščeglov, Y.K. The Novels by Ilf and Petrov. Readers's Companion; Ivan Limbach Publishing House: St. Petersburg, Ru ssia, 2009; 656p, ISBN 9785890591340.
34. Amlinski, I. 12 Chairs from Mikhail Bulgakov; Kirschner: Berlin, Germany, 2013; 328p, ISBN 9783000432842.
35. Zenkov, A.; Zenkov, E.; Belke, A. A Novel Text Analysis Method: Numerals Reveal the Author. In Proceedings of the Inte rnational Scientific Conference on New Industrialization and Digitalization (NID 2020), Ekaterinburg, Russia, 12 Decem ber 2020; EDP Sciences: Les Ulis, France, 2021; Volume 93, p. 03026.
36. Gan, G.; Ma, C.; Wu, J. Data Clustering: Theory, Algorithms, and Applications, ASA-SIAM Series on Statistics and Appli ed Probability; SIAM: Philadelphia, PA, USA, 2007.