Data Amplification Using the Iterated Newcomb-Benford Distribution

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

20-08-2021 / 25-08-2021

CITATION

Kak, Subhash (2021): Data Amplification Using the Iterated Newcomb-Benford Distribution. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.15832083.v1>

DOI

[10.36227/techrxiv.15832083.v1](https://doi.org/10.36227/techrxiv.15832083.v1)

Data Amplification Using the Iterated Newcomb-Benford Distribution

Subhash Kak

Chapman University, Orange CA, USA

and

Oklahoma State University, Stillwater, OK, USA

Email: subhash.kak@okstate.edu

Abstract

The Newcomb-Benford (NB) distribution of first digits has been applied widely in many areas ranging from engineering to natural and biological sciences for the investigation of self-similarity and randomness. In this article, we consider systems for which the data is not enough to obtain proper first digit statistics, and we propose the use of an iterated version of the distribution where the statistics are aggregated over different scales on grounds that the first digit distribution is approximately scale invariant across a wide range of phenomena and also because scaling and recomputing first digits is not a linear process and so this process generates new data. We provide examples of the use of the iterated test for data in two different biological applications, viz. that of the secretome and the genetic code in both of which the raw data does not include all the nine different first digits. The paper includes proposals for further research on the idea of data amplification using scaling transformations.

Keywords: Newcomb-Benford law, biological networks, self-similarity, connectivity, first digit phenomenon, genetic code

1. Introduction

The Newcomb-Benford (NB) Law [1-4], also called the first digit phenomenon, has been found to be approximately true in diverse phenomena ranging from networks, natural processes, geophysics, accounting and forensics [5-9]. It is also of much potential value in fields where randomness and fidelity of experimental design are of particular importance [10-12].

In one biological study, data on the number of open reading frames (ORFs) coded by genomes within prokaryota (unicellular organisms that lack a membrane-

enclosed nucleus) and eukaryota (organisms whose cells have a nucleus) shows that the NB distribution produces excellent fits to the prokaryote data and with modification also to the eukaryote data [11] consistent with an evolutionary development. It is most surprising that NB is nearly true for the dimensionless constants of nature (e.g. [6]) although the reason behind that is not known.

It is generally recommended that the sample size be at least 1000 to check whether data departs from the NB distribution [13], but this may not be achieved in many experimental situations. This is particularly true of biochemical networks -- such as metabolic networks, protein-protein networks, and genetic regulatory networks -- where structural constraints limit the connections and provide limited variety of data for analysis. One defining feature of such networks appears is self-similarity across different layers and the occurrence of fractal phenomena [14-16]. Since self-similarity is related to noninteger dimensionality [17-20], this provides another perspective on the problem which underlies our proposed method for overcoming the limitation on the volume of data.

This paper proposes that in constrained-connectivity networks the statistics can be enhanced by repeated use of scaling on the available data since it is known that the NB distribution is scale invariant and natural systems are approximately NB distributed. The fragments of the aggregated data are not linearly dependent due to the process of recomputation of the first digits subsequent to scaling and this, therefore, constitutes a method of data amplification. An application of this to the data of the secretome and the genetic code is presented and it is confirmed that this leads to superior statistics which is consistent with an evolutionary basis to the NB distribution.

2. Preliminaries

If a counting process is uniformly distributed over the range $\{1, \dots, S\}$, with random values of S , then the sum of a large number of these will satisfy the NB Law [3][4], where the leading digit n ($n \in \{1, \dots, r - 1\}$) for number to the base r , $r \geq 2$, occurs with probability as a logarithmic function:

$$P_{NB}(n) = \log_r(1 + \frac{1}{n}) \quad (1)$$

When the number consists of several digits, the same law applies with n replaced by the appropriate number. The NB law is scale invariant [3], and if numbers in the

data set are rescaled to another base, the probabilities become adjusted for the new base. For example, if numbers are represented to base 4, the first digit probabilities will be $P(1)=\log_4(2)=0.5$; $P(2)=\log_4(3/2)=0.292$; and $P(3)=\log_4(4/3)=0.2075$, and so half the random numbers will begin with the digit 1.

When comparing the NB distribution for numbers in different bases, the probability of a specific digit $n < r_1, r_2$, which are the two bases under consideration, is

$$\frac{P_{r_1}(n)}{P_{r_2}(n)} = \frac{\log(r_2)}{\log(r_1)} \quad (2)$$

Thus the digit frequencies for 1, 2, and 3 in base 4 are $\log(10)/\log(4)$ of the values for the same digits to base 10.

Let the NB probability distribution be represented as a vector $P(n): (p(1), p(2), \dots, p(r-1))$. We can speak of the companion vector $Q(n)$ for this distribution $Q(n) = (\sqrt{p(1)}, \sqrt{p(2)}, \dots, \sqrt{p(r-1)})$ as a unit vector in a $(r-1)$ dimensional space.

For $r=10$, that is base-10 distribution, the vectors $P_{NB}(n)$ and $Q_{NB}(n)$ are as below:

$$P_{NB}(n): (0.301, 0.176, 0.125, 0.097, 0.079, 0.067, 0.058, 0.051, 0.046)$$

$$Q_{NB}(n): (0.549, 0.420, 0.354, 0.311, 0.281, 0.259, 0.241, 0.226, 0.214)$$

(3)

Note that the first digit has 30% and the top 3 digits (one-third of the total) have just over 60% of the probability.

The norm of the vector $Q(n)$:

$$\|Q(n)\| = \sum Q_i^2 = \sum p_i = 1 \quad (4)$$

The closeness or nearness between two distributions $P_1(n)$ and $P_2(n)$ may be computed by measuring the distance d between the corresponding vectors $Q_1(n)$ and $Q_2(n)$.

The Euclidean distance $d(Q_1, Q_2)$ may be written in terms of the inner product for the vectors, for $Q_1, Q_2 \in \mathbb{R}^N$ are normalized, that is $\langle Q_1, Q_1 \rangle = \langle Q_2, Q_2 \rangle = 1$.

$$\begin{aligned} d(Q_1, Q_2)^2 &= \langle Q_1 - Q_2, Q_1 - Q_2 \rangle = \langle Q_1, Q_1 \rangle + \langle Q_2, Q_2 \rangle - 2\langle Q_1, Q_2 \rangle \\ &= 2(1 - \langle Q_1, Q_2 \rangle) \end{aligned} \quad (5)$$

We will define this distance to apply to the probability vectors P_1, P_2 and, therefore, we can write:

$$d(P_1, P_2) = (2(1 - \langle Q_1, Q_2 \rangle))^{1/2} \quad (6)$$

The maximum value of $\langle Q_1, Q_2 \rangle$ is 1 obtained for distributions that are identical and this corresponds to zero distance. The maximum distance between any two distributions will be when $\langle Q_1, Q_2 \rangle = 0$. This will correspond to:

$$d_{max}(P_1, P_2) = \sqrt{2} \quad (7)$$

In other words, the distance values will range from 0 (when they are identical) to $\sqrt{2}$ (when they are most dissimilar). The uniformly distribution P_U corresponds to:

$$P_U = \frac{1}{(r-1)}(1, 1, \dots, 1)$$

for which

$$Q_U = \frac{1}{\sqrt{r-1}}(1, 1, \dots, 1) \quad (8)$$

Theorem. The distance from any distribution to the uniform distributions is $d(P_U, P_A) = (2 - 2 \sum \sqrt{p_i} / \sqrt{(r-1)})^{1/2}$

Proof. The inner product $\langle Q_U, Q_A \rangle$ will be $\frac{1}{\sqrt{r-1}} \sum \sqrt{p_i}$, and the result follows from a simple substitution.

Example 1. Consider three distributions for base $r = 4$ with probabilities A: (0.6, 0.3, 0.1); B: (0.1, 0.2, 0.7); U: (0.333, 0.333, 0.333). We compute the corresponding Q vectors: Q_A : (0.775, 0.548, 0.316); Q_B : (0.316, 0.447, 0.837); Q_U : (0.577, 0.577, 0.577).

It is easily seen that $\langle Q_A, Q_B \rangle = 0.754$; $\langle Q_A, Q_U \rangle = 0.945$; $\langle Q_B, Q_U \rangle = 0.923$. Therefore, we get the table of distances as below:

Table 1. Mutual distances for vectors of Example 1

Distance, d	A	B	U
A	0	0.701	0.331
B	0.701	0	0.392
U	0.331	0.392	0

3. Relationship to the power law and natural data

NB is related to the power law [21] (probability is proportional to a^{-n} , for some appropriate constant a), and in particular to the Zipf's law according to which the rank-frequency distribution is an inverse relation [22][23]. This is seen in Table 2 that presents a comparison between NB and Zipf where for the latter the maximum frequency was matched to the leading NB frequency, which is necessary because unlike the limited leading digits of NB, the number of independent points of Zipf distribution can be extremely large. Table 2 is only showing the probabilities of the leading 9 elements in the dataset associated with the Zipf distribution (such as the set of words in a dictionary).

Table 2. First digit probabilities for base 10 and Zipf's probabilities adjusted for peak at 0.301

	1	2	3	4	5	6	7	8	9
1st digit	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046
Zipf's freq	0.301	0.155	0.103	0.075	0.060	0.051	0.043	0.038	0.034

The mean and the standard deviation of the NB distribution are:

$$\mu_{NB} \cong 3.441; \sigma_{NB} \cong 2.462 \quad (9)$$

These will be used for comparison later with the distribution of different real data.

Figure 1. Blue line (NB law for $r = 17$); green line (Zipf's law)

Figure 2 presents a more general power law $0.3 \times a^{-n}$, that shows how one can come closer to the NB law by appropriate choice of the constant.

Figure 2. Power law $0.3 \times a^{-n}$, where a takes values 1.4 (red), 1.5 (blue), 1.7 (green), 1.8 (purple)

The observational data from a variety of natural phenomena in Table 3, which is adapted from [6], shows that the data appears to largely conform to the frequencies expected from the NB Law (the top three are almost always at least 60% of the total) but with small deviations. One can assume that these deviations have the potential to offer further insight into the nature of the data and the underlying physical process.

Table 3. First digit data from a variety of natural phenomena, abridged from [6]

<i>First Digit Frequencies percentage</i>									
	1	2	3	4	5	6	7	8	9
NB's Law	30.1	17.6	12.5	9.69	7.92	6.69	5.80	5.12	4.58
Fund. Phys. constants	34.0	18.4	9.2	8.28	8.58	7.36	3.37	5.21	5.52
Geomagnetic field	28.9	17.7	13.3	9.4	8.1	6.9	6.1	5.1	4.5
Geomagnetic Reversals	32.3	19.4	13.9	11.8	5.3	4.3	3.2	5.4	4.3

Fermi Space Teles. fluxes	30.3	17.9	13.0	9.9	7.6	6.96	5.23	5.23	2.72
Pulsars rotation freq	33.9	20.7	12.7	7.6	5.3	5.0	4.94	4.67	4.88

The significance of the data of Table 3 is that one can expect similar near-NB distribution to be valid for different phenomena in natural systems and, therefore, one can justifiably use scaling transformation for data amplification when the data is not available in sufficient amount. It also follows that the nature of the distributions is a consequence of some evolutionary mechanism underlying the physics.

4. The Iterated NB (INB) distribution

In most systems, the data can be investigated directly to determine the relevant NB digit probabilities. But what if the data set is small and not even all the digits are represented? This situation can arise where the collection of the data beyond the meagre set is difficult and expensive, or where physical constraints make only certain kinds of interactions possible.

We saw that the first digit probabilities $P(n)$, ($n \in \{1, \dots, 9\}$) constitute a vector in an appropriate 9-dimensional space, with non-zero projections on each of the 9 bases.

$$P(n) = P(n_i), i = 1, \dots, 9 \quad (10)$$

Assuming that the data is being processed as it comes in, the computed vector will move closer to the NB point and if the data is sufficiently large, it will settle there, since it is known that the NB probabilities do not change as the scale is changed [3]. The scaling is accomplished by multiplying all components of the vector by some scaling constant and then recomputing the first digit probabilities. It is the recomputation of the digit frequencies that renders the system non-linear.

Let the computed outcomes with respect to the first digit in the constrained data situation be called $S(n)$, where some of the values can be zero if the data is limited. In order to populate these zero components, we will multiply the vector $S(n)$ with different scaling factors to generate a full vector that we call the aggregated data vector $S'(n)$.

The fact that we propose the use of only a certain number of scaled terms may appear somewhat arbitrary and, in principle, the scaling can be done as often as it

does not lead to a duplication but that possibility that will not be investigated in the present paper.

Example 2. Let $S(n) = (1, 1, 3, 9)$. By itself, it has probabilities $P(n) = (0.5, 0, 0.25, 0, 0, 0, 0, 0, 0.25)$, with mean, μ , 3.5 and s.d., σ , of 3.279. It has three independent non-zero values, and therefore we need to multiply with 6 different constants to generate the missing values in the vector.

$$Q_S(n): (0.707, 0, 0.5, 0, 0, 0, 0, 0, 0.5)$$

$$\langle Q_S, Q_{NB} \rangle = 0.707 \times 0.549 + 0.5 \times 0.354 + 0.5 \times 0.214 = 0.672$$

We generate the following matrix for scaled $S(n)$ using different multipliers b :

$$\begin{aligned} &1, 1, 3, 9 \\ &2, 2, 6, 18 (\times 2) \\ &4, 4, 12, 36 (\times 4) \text{ (} b=3 \text{ not used for it is in the original data)} \\ &5, 5, 15, 45 (\times 5) \\ &6, 6, 18, 54 (\times 6) \\ &7, 7, 21, 63 (\times 7) \\ &8, 8, 24, 72 (\times 8) \end{aligned}$$

This completes the six additional scaled values of the data; $b=9$ is not used for it is present in the original data.

The aggregated data is now:

$$S'(n): (1, 1, 18, 12, 15, 18; 2, 2, 21, 24; 3, 30; 4, 4; 5, 5, 54; 6, 6, 6, 63; 7, 7, 72; 8, 8; 9)$$

$$P'(n) : (6/28, 4/28, 2/28; 3/28; 3/28; 4/28; 3/28; 2/28; 1/28)$$

$$Q'_S(n) : (0.463, 0.378, 0.267, 0.327, 0.327, 0.378, 0.327, 0.267, 0.189)$$

$$\langle Q'_S, Q_{NB} \rangle = 0.978$$

Notice that the correlation of the data with the NB distribution has increased considerably (from 0.672 to 0.978 which is nearly 46% increase). The mean, μ , has increased from 3.5 to 4.179 and s.d., σ , has reduced from 3.279 to 2.494.

The application of the INB to real data will now be explained by means of two examples.

5. The human secretome

The secretome is the set of proteins expressed by an organism and secreted into the extracellular space. In humans, this subset of the proteome encompasses 13-20% of all proteins, including cytokines, growth factors, extracellular matrix proteins and regulators, and shed receptors. The secretome of a specific tissue can be measured by mass spectrometry and its analysis constitutes a type of proteomics known as secretomics [24][25].

As we can see from Figure 3, the data in the human secretome consists of number of proteins that are secreted to blood, brain, digestive system and so on is (the specific recipients are not important, the distribution of the numbers is):

$$S(n): 128, 173, 234, 20, 41, 65, 88, 729, 933$$

These numbers are written according the increasing first digit. There are no numbers that start with 3 or 5. The probabilities computed for the digits are:

$$P_S(n): 2/9, 2/9, 0, 1/9, 0, 1/9, 1/9, 1/9, 1/9$$

with mean, μ , 4.44 and s.d., σ , of 2.948.

We create additional two rows in our matrix by multiplying by 3 and 5. We have, in total:

$$\begin{array}{l} 128, 173, 234, 20, 41, 65, 88, 729, 933 \\ 384, 519, 702, 60, 123, 195, 264, 2187, 2799 \\ 640, 865, 1170, 100, 205, 325, 440, 3645, 4665 \end{array}$$

$$P'(n): (6/27, 6/27, 3/27, 3/27, 1/27, 3/27, 2/27, 2/27, 1/27)$$

Figure 3. The human secretome with numbers associated with the named classes

The mean of the amplified data set is 3.74 and the standard deviation is 2.488 which are closer to the corresponding NB values (3.44 and 2.462) than the raw data above (4.44 and 2.948).

6. The genetic code

The randomness of the standard genetic code [26] as related to the power law (or Zipf's law) was investigated recently [27]. It should be pointed out that the origin of the triplet code is being investigated from several different perspectives of optimization [28][29]. The previous study, therefore, may be seen as belonging to this larger class of studies. The specific question addressed in that paper was the logic behind the non-uniformity of the mapping of the codons into amino acids.

Table 4. Number of codons to amino acid

Type name	A	B	C	D	F	G
codons to amino acid	1	2	3	4	5	6
number of this type	2	9	1	5	0	3

The assignment of codons ranges from 1 to 6 as shown in the graph below:

Figure 4. Codons to amino acids as different groups

The number of each type is also in a broad distribution with values ranging from 0 to 9, with the rank order frequencies of 9, 5, 3, 2, 1, 0 which we have shown earlier is quite consistent with the Zipf distribution. In other words, the natural data did correspond to the optimality principle that generates this particular distribution.

Since all information spaces, including those associated with cognitive processes, will have a similar optimality structure, it is quite natural to see the Zipf law show up both in codon frequencies as well as word frequencies of a natural language. We will now check it against the NB distribution.

The six connectivity numbers (0, 1, 2, 3, 5, 9) give us information on only 5 different first digits (it is missing four digits), with mean of 4 and standard deviation of 2.828.

The numbers can be rewritten as below by multiplying by 4, 6, 7, and 8 to give us the following array:

1	2	3	5	9
4	8	12	20	36
6	12	18	30	54
7	14	21	35	63
8	16	24	40	72

$P_S(n)$: (6/25, 4/25, 4/25, 2/25, 2/25, 2/25, 2/25, 2/25, 1/25), as shown in Figure 5.

Figure 5. The distribution $P_S(n)$ for the genetic code

This roughly approximates the NB distribution. Its mean and standard deviation are 3.8 and 2.51, and both these values are closer to the NB values of 3.44 and 2.46.

Displacing and then scaling

The original data set had 6 numbers out of which one was zero and invariant with respect to scaling multiplication. We now add 1 to each data point (1 2 3 5 9 0) that gives us (1 2 3 4 6 10) and then subsequently perform the scaling operations. The first row is the enhanced numbers and the rows below are multiplication by scaling factors of 5, 7, 8, 9, and so on:

1	2	3	4	6	10
5	10	15	20	30	50
7	14	21	28	42	70
8	16	24	32	48	80
9	18	27	36	54	90

Aggregating, we get the following probabilities for $P'_S(n)$:

$$P'_S(n): (7/30, 6/30, 4/30, 3/30, 3/30, 1/30, 2/30, 2/30, 2/30)$$

Figure 6. The distribution $P'_S(n)$ for the genetic code

For this case, the mean is 3.733 and the standard deviation is 2.555.

Table 5 summarizes the mean and standard deviation values for the raw data, the amplified data with 5 points, and the amplified data obtained by translation followed by scaling.

	mean	standard deviation
raw data	4	2.828
amplified data (scaling)	3.8	2.514

Amplified data (translation and scaling)	3.73	2.555
NB	3.44	2.46

Clearly the amplified data statistics are closer to the underlying assumed near-NB distribution than was the case for the raw data.

7. Discussion

In this article, we considered structured systems for which the data is not enough to obtain proper NB statistics on first digits, and we proposed the use of an iterated version of the distribution where the statistics are aggregated over different scales on grounds that the first digit distribution is approximately scale invariant across a wide range of phenomena and also because scaling and recomputing first digits is not a linear process and so this process amplifies data. It was also shown that if the raw data has unusable zeros then all the data can be translated prior to the scaling operation.

We provided example of the use of the iterated test for data in two different biological applications, viz. that of the secretome and the genetic code. In each of these cases, we show that the use of scaling (and translation) increases the power of the data by bringing it closer to the putative NB distribution.

Clearly this method can be used in applications from a variety of fields including accounting and forensics.

The process of data amplification suggested in this paper needs to be further investigated since our method of adding scaled versions of the data to fill in the missing first digits does not address the question of the optimality of the amount of scaled data that should be added.

References

1. Newcomb, S. Note on the frequency of use of different digits in natural numbers, *Am. J. Math.*, 4,39–40 (1881)
2. Benford, F. The law of anomalous numbers, *Proc. Am Philos. Soc.*, 78(4), 551–572 (1938)
3. Hill, T. P. A statistical derivation of the significant-digit law, *Statistical Science* 10, 354-363 (1995)
4. Hill, T. P. (1998), The first-digit phenomenon, *Am. Sci.*, 86, 358–363 (1998)

5. Kak, S. Variations of the Newcomb-Benford law. arXiv (2018).
<https://arxiv.org/ftp/arxiv/papers/1806/1806.06695.pdf>
6. Sambridge, M., H. Tkalić, and A. Jackson (2010), Benford's law in the natural sciences, *Geophys. Res. Lett.*, 37, L22301
7. Prandl, S. et al (2017), An Investigation of Power Law Probability Distributions for Network Anomaly Detection, *IEEE Security and Privacy Workshops (SPW)*, 217-222.
8. Nigrini, M. J., and S. J. Miller (2007), Benford's law applied to hydrology data—Results and relevance to other geophysical data, *Math. Geol.*, 39, 469–490.
9. Kak, S. Power series models of self-similarity in social networks. *Information Sciences*, vol. 376, pp. 31-38, 2017.
10. McCoy, M.M., Allen, A.P., Gilloly, J.F. The random nature of genome architecture: predicting open reading frame distributions. *PLoS ONE* 4(7): e6456.
11. Friar J.L., Goldman T., Perez–Mercader J. (2012) Genome sizes and the Benford distribution. *PLoS ONE* 7(5), e36624.
12. Kak, S. Significance testing in natural and biological systems: a review. *TechRxiv* (2021). <https://doi.org/10.36227/techrxiv.14599836.v1>
13. Wong, S.C.Y. Testing Benford's Law with the First Two Significant Digits. Thesis, University of Victoria (2010)
14. Mandelbrot, B. B., *The Fractal Geometry of Nature*. W. H. Freeman (1983)
15. Falconer, K.J., *Fractal Geometry: Mathematical Foundations and Applications*. (Wiley, 2003)
16. Kak, S. Information theory and dimensionality of space. *Scientific Reports* 10, 20733 (2020).
17. Kak, S. The intrinsic dimensionality of data. *Circuits Syst. Signal Process.* 40, 2599–2607 (2021); <https://doi.org/10.1007/s00034-020-01583-8>
18. Kak, S. Fractals with optimum information dimension. *Circuits Syst. Signal Process.* 40 (2021); <https://link.springer.com/article/10.1007/s00034-021-01726-5>
19. Kak, S. The base-e representation of numbers and the power law. *Circuits Syst. Signal Process.* 40, 490–500 (2021)
20. Kak, S. Asymptotic freedom in noninteger spaces. *Scientific Reports* 11, 1–5 (2021).
21. Newman, M.E.J. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46, 323-351 (2005)
22. Vicsek, T. *Fluctuations and scaling in biology*. Oxford University Press (2001)
23. Zipf, G.K. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Reading, MA (1949)
24. Mukherjee, P., Mani, S. Methodologies to decipher the cell secretome. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics.* 1834 (11), 2226–32 (2013)
25. Uhlén M et al. Proteomics. Tissue-based map of the human proteome. *Science.* 347 (6220), 1260419 (2015)
26. Kak, S. The e-dimensionality of genetic information. *TechRxiv* (2021).
<https://doi.org/10.36227/techrxiv.14977479.v1>

27. Crick, F.H.C. The origin of the genetic code. *J. Mol. Biol.* 38, 367–379 (1982)
28. Baranov, P.V., Venin, M., Provan, G. Codon size reduction as the origin of the triplet genetic code. *PLoS ONE* 4(5): e5708. (2009)
29. Attie, O., Sulkow, B., Di, C., Qiu, W. Genetic codes optimized as a traveling salesman problem. *PLoS ONE* 14(10): e0224552. (2019);
<https://doi.org/10.1371/journal.pone.0224552>