Journal of Statistical Modelling: Theory and Applications
Vol. 6, No. 1, 2025, pp. 49-57
Yazd University Press 2025

JSMTA

Research Paper

Double-crossing Benford's law

Javad Kazemitabar*
Department of Electrical and Computer Engineering, Babol Noshirvani University of Technology, Babol, Mazandaran, Iran

Received: October 19, 2024/ Revised: September 09, 2025/ Accepted: September 12, 2025

Abstract: From Covid-19 mortality rate to image tampering, Benford's law is used to detect fraudulent activities. The underlying assumption for using the law is that a "regular" dataset follows the significant digit phenomenon. In this paper, we address the scenario where a shrewd fraudster manipulates a list of numbers in such a way that while providing the desired statistics, it still complies with Benford's law. We develop a framework that offers several degrees of freedom to such a fraudster, such as the minimum, maximum, mean, and size of the manipulated dataset. The conclusion further corroborates the idea that Benford's law -if at all- should be used with utmost discretion as a means for fraud detection.

Keywords: Benford's law; Distribution; Forensic; Fraud detection; Statistical analysis.
Mathematics Subject Classification (2010): 62P05.

1 Introduction

Ever since Benford's law was suggested as a test of *naturalness*Benford (1938), researchers have explored many areas to apply the law. Examples include-but is by no means limited to-checking national Covid-19 mortality rate Sambridge and Jackson (2020), digital image tampering Parnaket al. (2022), social welfare fraudda Silva Azevedo et al. (2021), tax Nigrini (1996), and financial statements fraud Zack (2013). Benford's law has also been used to investigate Natural Hazard dataset homogeneity Joannes-Boyau et al. (2015) or inflation data at governmental level Miranda-Zanettia

*Corresponding author: j.kazemitabar@nit.ac.ir

et al. (2019). Many articles, books and other resources related to Benford's law, including theoretical and applied, could be seen in "Benford Online Bibliography" Berger et al. (2009).

Deviation of a list of numbers from Benford's law is usually considered as a red flag. The interested reader is referred to Miller (2015) and Nigrini (2020) for a detailed explanation of Benford test application and different methods of harnessing the Benford test to find anomaly in data. It has been suggested, however, that perfect adherence to this law could also imply manipulation as we expect small levels of deviation from the law in regular lists of numbers Kalinin and Mebane (2017). A question is then raised as whether it is possible to systematically manipulate a list such that it still complies with the law. In this paper we address this question, but let us first review some of the efforts in scientifically explaining Benford compliant distributions.

The rest of this paper is organized as follows. Section 2 reviews Benford compliant distributions. Section 3 covers construction of fake data. In Section 4, we discuss the results of two examples. Finally, Section 5 concludes the paper.

2 Benford-compliant distributions

Hill's 1995 paper Hill (1995) provides a statistical explanation of Benford's law. The author shows that "if probability distributions are selected at random, and random samples are then taken from each of these distributions in any way so that the overall process is scaled(or base) neutral" then Benford's law holds. He then asks "An interesting open problem is to determine which common distributions (or mixtures thereof) satisfy Benford's law". Several researchers pursued this question and found conditions for a Benford compliant distribution Balanzario and Sánchez-Ortiz (2010); Balanzario (2015); Leemis et al. (2000); Berger and Hill (2015). They also proposed example distributions that satisfy conditions mentioned above. However, the proposed distributions do not provide the necessary degrees of freedom for a fraudster to build synthetic * Benford compliant samples with desired statistics. Similar effort was presented in Haracci and Haracci (n.d.) where a synthetic Benford dataset is generated using a software. However, to the best of our knowledge, there has been no report of systematic methods producing customized Benford compliant datasets. Based on a recent report Kazemitabar (2023), in this paper, we provide two families of Benford-complaint distributions with tunable parameters providing us with several degrees of freedom such as minimum, maximum, mean and size, to generate such synthetic datasets. It should be noted however, that we are by no means encouraging fraudsters to use these algorithms. Our goal is solely to show that it can be performed and that the auditors should be careful not to rely too much on these tests. The existence of such distributions shows that Benford's law should be carefully used as a means of fraud detection.

In Leemis et al. (2000), a few Benford compliant distributions were proposed that are the building blocks of the distributions to be introduced in this paper.

• *Example 1:* Let $Y \sim U(0, 2)$. Then, $X = 10^Y$ is a Benford compliant distribution defined in $(10^0, 10^2)$. The result can be generalized for $Y \sim U(a, b)$ for integer a and b.

*The term *synthetic* Benford set was first used by the celebrated author Mark Nigrini Nigrini (2020). He provides a method based on the uniform mantissa concept to build synthetic Benford compliant samples, where the user can designate the maximum and minimum of the generated numbers.

• *Example 2:* Let $Y \sim Triangular(0, 1, 2)$. In other words,

$$f_Y(y) = \begin{cases} y, & 0 < y < 1, \\ 2 - y, & 1 \leq y < 2. \end{cases}$$

then, $X = 10^Y$ is a Benford compliant distribution defined in $(10^0, 10^2)$. The result can be generalized to symmetric Triangular distributions of Y such as $Triangular(a, b, c)$ where a, b, and c are all integers and $b = (a + c)/2$.

In both of the above examples, even though the maximum and minimum of the distribution -in its general form- is tunable, the average is not. To amend this shortcoming we use the a lemma that was independently proven by a number of authors Kazemitabar and Kazemitabar (2020)Balanzario and Sánchez-Ortiz (2010)Balanzario (2015).

Lemma 2.1. *If* $\sum_{k=-\infty}^{+\infty} f_Y(z + k) = 1$ *then,* $X = 10^Y$ *is a Benford compliant distribution.*

Using this lemma, we build upon these examples to introduce our tunable distributions. Concretely, we design the distributions such that the shifted versions of the density function add up to 1.

Distribution design procedure: To come up with the following two distributions, a few things were taken into consideration. First and foremost, in order to be able to tune the mean of the distribution, one needed to introduce an extra parameter. This parameter which is represented by "a" in both of these distributions, ranges from slightly more than zero to infinity. The role of this parameter is to shift the mean. One knows that the mean of any distribution is limited between the minimum and maximum value that distribution takes. In our proposed distribution, we were able to shift the mean towards these two extremes using very small and very large values for "a". To be precise, for very small values of "a", we were able to fill in the lion share of the distribution near the minimum causing the mean to be close to it. Similarly, by choosing large values for "a" we will shift the major part of the area under curve of the density function towards the maximum value. The two proposed distributions are different in the amount of "shift" towards maximum and minimum. The first distribution is able to get closer to both the minimum and the maximum. The advantage of the second distribution, can be described as being "less obvious" as a synthetic dataset, due to its non-uniform shape in the intervals.

• First Proposed Distribution

If

$$f_{Y_1}(y) = \begin{cases} \frac{a}{\sum_{t=1}^{t=K} a^t}, & m \leq y < m + 1, \\ \frac{a^2}{\sum_{t=1}^{t=K} a^t}, & m + 1 \leq y < m + 2, \\ \vdots & \\ \frac{a^K}{\sum_{t=1}^{t=K} a^t}, & m + K - 1 \leq y < m + K, \end{cases}$$

then, $X_1 = 10^{Y_1}$ is a Benford compliant distribution with the following statistics:

$$\min(X) \quad = \quad 10^m$$

$$\begin{aligned} \max(X) &= 10^{m+K} \\ mean(X) &= \frac{9a}{Ln(10).\sum_{t=1}^{t=K} a^t}.10^m.\frac{1-(10a)^K}{1-10a}, \end{aligned} \tag{1}$$

where mean(X_1) ranges between 3.9×10^m and $3.9 \times 10^{m+K-1}$ for very small and very large values of a respectively.

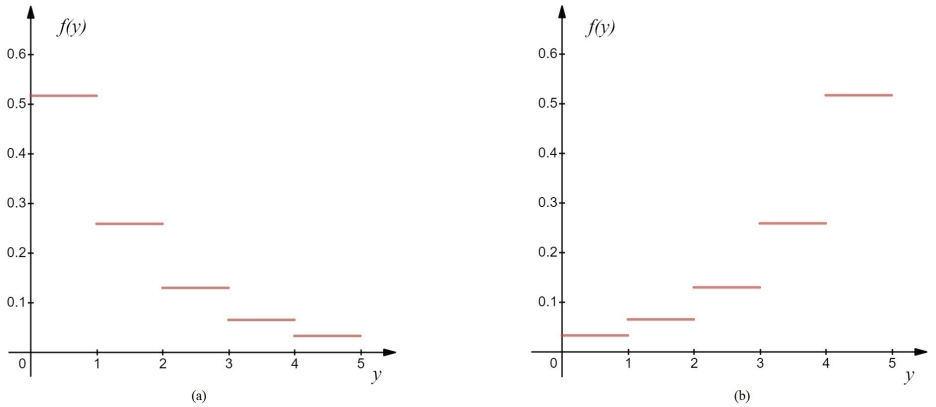

Figure 1: An example Y_1 distribution with $m = 0$, $k = 5$ and (a) $a = 0.5$ (b) $a = 2$. $X_1 = 10^{Y_1}$ follows Benford's law.

• Second Proposed Distribution
If

$$f_{Y_2}(y) = \begin{cases} \frac{a}{\sum_{t=1}^{t=K} a^t}(y-m), & m \le y < m+1, \\ \frac{a}{\sum_{t=1}^{t=K} a^t} - \frac{a}{\sum_{t=1}^{t=K} a^t}(y-m-1), & m+1 \le y < m+2, \\ \frac{a^2}{\sum_{t=1}^{t=K} a^t}(y-m-2), & m+2 \le y < m+3, \\ \frac{a^2}{\sum_{t=1}^{t=K} a^t} - \frac{a}{\sum_{t=1}^{t=K} a^t}(y-m-3), & m+3 \le y < m+4, \\ & \vdots \\ \frac{a^K}{\sum_{t=1}^{t=K} a^t}(y-m-2K+2), & m+2K-2 \le y < m+2K-1, \\ \frac{a^K}{\sum_{t=1}^{t=K} a^t} - \frac{a}{\sum_{t=1}^{t=K} a^t}(y-m-2K+1) & m+2K-1 \le y < m+2K, \end{cases}$$

then, $X_2 = 10^{Y_2}$ is a Benford compliant distribution with the following statistics:

$$\begin{aligned} \min(X_2) &= 10^m, \\ \max(X_2) &= 10^{m+2K}, \\ mean(X_2) &= \frac{99-81/Ln(10)}{Ln(10).\sum_{t=1}^{t=K} a^t}.10^m.a.\frac{1-(100a)^K}{1-100a}, \end{aligned}$$

where mean(X_2) ranges between $2.7 \times 10^{m+1}$ and $2.7 \times 10^{m+2K-1}$ for very small and very large values of a respectively.

One might wonder if the maximum and minimum points in the above mentioned distributions have to be powers of 10. To answer this, we should recall that compliance

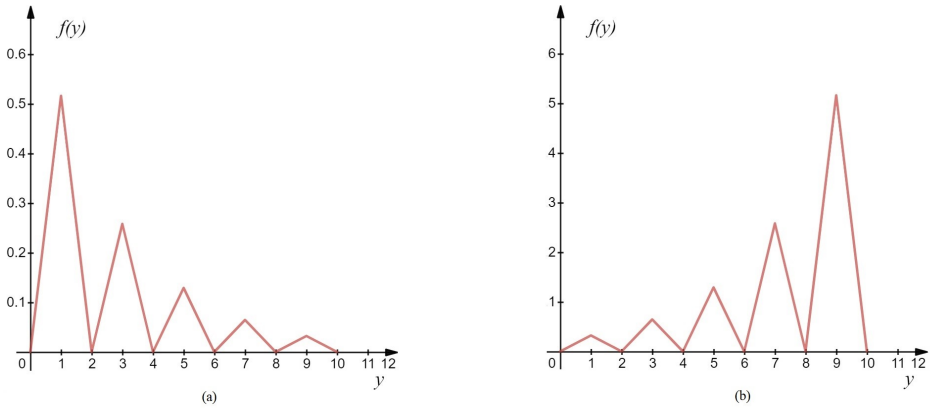

Figure 2: An example Y_2 distribution with $m = 0$, $k = 5$ and (a) $a = 0.5$ (b) $a = 2$. $X_2 = 10^{Y_2}$ follows Benford's law.

with Benford's law is scale-invariant. As such the generated numbers can be multiplied with a constant number. Nevertheless, both the above proposed distributions require that the max and min are apart by an integer power of 10, that is $\frac{max}{min} = 10^K$ in the first distribution and $\frac{max}{min} = 10^{2K}$ in the second.

3 Constructing fake data

In this section, we show how a fraudster can generate a Benford compliant dataset. We provide two examples. The first example is about journal entries of a company trying to look profitable and the second example is on fake mortality rates. Of course, in order to fake a journal entry, the fraudster needs to generate two separate datasets; one for income and the other for expenses. For each dataset We can tune the maximum and minimum as well as the number of items and the total sum. This is directly achieved by plugging the right value for m, K and a in the distributions introduced in the previous section. Moreover, we note that total sum of numbers in the dataset is equal to the size of that dataset multiplied by its average. Since, we have control over size and average, as a result we have control over total sum. We use *inverse transform sampling* Luc (1986) to generate random samples. The goal that the inverse transform sampling technique achieves can be summarized is as follows:

- Let Y be a random variable with cumulative distribution function F_Y.
- We want to generate samples of Y according to the given distribution.

To do so, the inverse transform sampling technique first, generates uniform samples in the interval $[0, 1]$. Next, it finds the inverse of the given CDF, i.e. $F_Y^{-1}(u)$. Finally, the technique calculates $Y'(u) = F_Y^{-1}(u)$. The resulting random variable $Y'(U)$ will have the desired distribution represented by the given CDF, i.e. F_Y. That is simply because $\Pr(F_Y^{-1}(u) \leq y) = \Pr(U \leq F_Y(y)) = F_Y(y)$.

3.1 Example 1

Suppose a hypothetical company's income and expenses each total 5700000 $ and 2310000 $ respectively. Also, let us assume there are 1320 income entries in the journal ranging from 1000$ to 100000$ and 760 expense related entries in the range of 100$ to 100000$. Using (1), we find m, K and a to be 3, 2, and 0.01177886831 respectively for income related entries. Moreover, for expense related entries, we find the aforementioned parameters to be 2, 3, and 0.25927727232382797. While for the scenario at hand we were able to analytically solve for a, in general, however, numerical methods may be necessary specially when K is a large number. Figure 3 shows the histograms of income and expense entries. We then generate $X = 10^Y$ to populate the journal entries for revenue and expense separately. The total sum for revenue samples add up to 5556356 which is 97% accurate compared to the requested revenue of 5700000$. As for the expense dataset, the sum of fake journal entries is 2192381 which shows 5% deviation from the desired expense total of 2310000$.

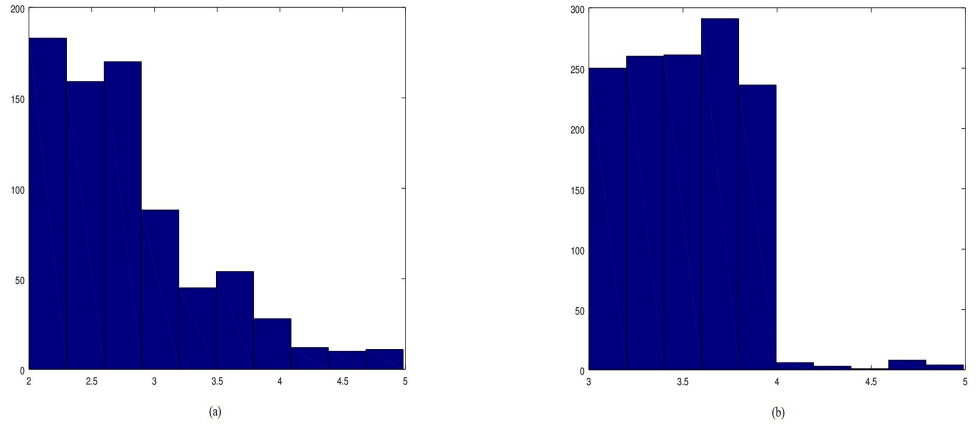

Figure 3: Histogram of the synthetic (fake) data generated based on the proposed Y_1 distribution. The actual journal entries will be populated by taking 10 to the power of these numbers. (a) Expense related Y samples with $m = 2$ and $K = 3$ (b) Income related Y samples with $m = 3$ and $K = 2$.

3.2 Example 2

Suppose a fraudster is generating covid-related mortality rate. To cover up for large number of victims, the health ministry decides to generate (fake) lower numbers (One cannot emphasize enough that this example is a completely hypothetical scenario). The number list includes death rate from mid March 2020 for 110 consecutive days. The ministry is ordered by authorities to limit the daily mortality rate between 13 to 1005 with an average of 56. To *cook the books*, the ministry will use the first proposed distribution with $m = 1$ and $K = 2$ to satisfy the max and min requirements. From the last line in (1) one can calculate $a = 0.0505085$ to make sure the daily average mortality rate stays around the desired value. Using these parameters, we applied the inverse transform sampling technique on 110 samples of uniformly distributed numbers

to obtain the mortality list. Since the mortality rates need to be integer values, we then took the integer part of 10^Y to generate the final list. We got an average of 61, minimum of 10, and a maximum of 997 which were close fits from the corresponding desired values.

4 Discussion

We tested the generated journal entries across 3 popular Benford tests namely chi-square, mantissa-arc and mean absolute deviation (MAD). These three tools are among the toughest tests that auditors apply for Benford tests. Often times, even untampered sets with large samples fail to meet the requirements of these tests Nigrini (2020). Yet, the generated fake dataset in our example, was able to pass. The results of all three tests are shown in Tables 1-2. To perform Benford tests, we used R package titled *Benford.analysis*. This package makes it very easy to perform chi-square, mantissa-arc and MAD tests on the data. The practice of generating fake Benford compliant datasets can easily be performed so long as the average is not too close to either end, i.e. minimum or maximum of the desired set. Concretely, the first proposed distribution, X_1, ranges between 3.9 times the minimum value, i.e. 10^m, and 0.39 of the maximum value, i.e. 10^{m+K}. In practical scenarios, it rarely happens that the dataset is skewed to the level that the average exceeds the aforementioned limits. As such, building fake data to deceive the auditor is usually achievable and thus the auditor shall not solely rely on Benford test.

Table 1: Benford test results confirm compliance of fake data for Example 1.

	Chi-squared test p-value	Mantissa arc test p-value	MAD
Revenue	0.9	0.93	Close conformity
Expense	0.54	0.28	Acceptable conformity

Table 2: Benford test results confirm compliance of fake data for Example 2.

	Chi-squared test p-value	Mantissa arc test p-value	Average Z-score
Mortality	0.57	0.83	1.18

5 Conclusion

Benford's law is used for detecting fraudulent reports in many fields. The underlying assumption for using the law is that a "regular" dataset follows the significant digit phenomenon. In this paper, we addressed the scenario where a shrewd fraudster manipulates a list of numbers in such a way that while providing the desired statistics still complies with Benford's law. We developed a framework that offers several degrees of freedom to such a fraudster such as minimum, maximum, mean and size of the manipulated dataset. The conclusion further corroborates the idea that Benford's law -if at all- should be used with utmost discretion as a means for fraud detection.

Acknowledgment

The author would like to thank Professor Reza Zahabi for his helpful comments.

References

Berger, A., Hill, T.P. and Rogers, E. (2009). *Benford Online Bibliography*. http://www.benfordonline.net.

Balanzario, E.P. (2015). Benford's law for mixtures. *Communications in Statistics-Theory and Methods*, **44**(4):698–709.

Balanzario, E.P. and Sánchez-Ortiz, J. (2010). Sufficient conditions for Benford's law. *Statistics & Probability Letters*, **80**(23-24):1713–1719.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, **78**:551–572.

Berger, A. and Hill, T.P. (2015). *An Introduction to Benford's Law*. Section 4.3, Princeton University Press.

da Silva Azevedo, C., Goncalves, R.F., Gava, V.L. and de Mesquita Spinola, M. (2021). A Benford's Law based methodology for fraud detection in social welfare programs: Bolsa Familia analysis. *Physica A: Statistical Mechanics and its Applications*, **567**:125626.

Haracci, M.O. and Haracci, G.B. (n.d.). *Benford Wizard*. https://www.members.tripod.com/benfordwiz/.

Hill, T.P. (1995). A Statistical derivation of the significant-digit law. *Statistical Science*, **10**(4):354–363.

Joannes-Boyau, R., Bodin, T., Scheffers, A., Sambridge, M. and May, S.M. (2015). Using Benford's law to investigate natural hazard dataset homogeneity. *Scientific Reports*, **5**(1):12046.

Kalinin, K. and Mebane Jr, W.R. (2017). When the Russians fake their election results, they may be giving us the statistical finger. *The Washington Post*, **11**.

Kazemitabar, J. (2023). A general framework for constructing distributions satisfying Benford's law. *Communications in Statistics-Simulation and Computation*, **52**(12):6160–6167.

Kazemitabar, J. and Kazemitabar, J. (2020). Measuring the conformity of distributions to Benford's law. *Communications in Statistics-Theory and Methods*, **49**(14):3530–3536.

Leemis, L.M., Schmeiser, B.W. and Evans, D.L. (2000). Survival Distributions Satisfying Benford's Law. *American Statistician*, **54**(4):236–241.

Luc, D. (1986). *Non-Uniform Random Variate Generation*, New York: Springer-Verlag.

Miranda-Zanetti, M., Delbianco, F. and Tohmé, F. (2019). Tampering with inflation data: A Benford law-based analysis of national statistics in Argentina. *Physica A: Statistical Mechanics and its Applications*, **525**:761–770.

Miller, S.J. (2015). *Benford's Law: Theory and Applications*, Princeton University Press.

Nigrini, M.J. (1996). A taxpayer compliance application of Benford's law. *The Journal of the American Taxation Association*, **18**(1):72.

Nigrini, M.J. (2020). *Forensic Analytics, Methods and Techniques for Forensic Accounting Investigations*. 2nd Edition, John Wiley & Sons.

Parnak, A., Baleghi, Y. and Kazemitabar, J. (2022). A novel image splicing detection algorithm based on generalized and traditional Benford's law. *International Journal of Engineering, Transactions A: Basics*, **35**(04):626–634.

Sambridge, M., and Jackson, A. (2020). National COVID numbers–Benford's law looks for errors. *Nature*, **581**(7809):384–385.

Zack, G.M. (2013). *Financial Statement Fraud Strategies for Detection and Investigation*. Chapter 18, John Wiley & Sons.