Benford's law

A heuristic derivation, simulation and explanation of the Newcomb-Benford law

Daniel Zahnd

May 14, 2024 - July 5, 2024

Abstract

This paper is concerned with a heuristic derivation of the Benford law from first principles. It is shown in the paper, that the NBL arises from the requirement of scale invariance alone. Scale invariance means, that the mantissa x of a number from a dataset has the same probability $p(x)\,\mathrm{d}x$ of being in the interval $[x, x+\mathrm{d}x]$, as λx has the probability $p(\lambda x)\,\mathrm{d}(\lambda x)$ of being in the interval $[\lambda x, \lambda x+\mathrm{d}(\lambda x)]$ for $\lambda \in \mathbb{R}$. The possible range for x or λx respectively is the interval $[1, 10]$.

The derived NBL is applied to a simulated dataset following the the Benford distribution, aswell as to various datasets of different nature.

1 Introduction

The Benford law was actually discovered by [Newcomb, 1881]. He observed, that the pages in logarithm books containing tables for logarithms with number one were much dirtier than the pages with logarithms of higher numbers. He proposed that the first digit 1 is much more likely to occur in a dataset spanning many orders of magnitude than a higher number larger than one. [Benford, 1938] rediscovered this law later on and formalized it. Only through Theodore Hill however, the Benford law came to be known by a greater mathematical community, because he applied Benford's distribution to solve real world problems.

Nowadays, Benford's law is used to detect frauds in various kinds of datasets. What is important to note here is that the law is only applicable to datasets spanning many orders of magnitude. The reason for this will become more clear with a heuristic derivation given below.

2 Derivation of the Newcomb-Benford law

Following [Burgos and Santos, 2021], Benford's law can be derived from scale invariance. Let $p(x)$ denote the probability density for the mantissa x of some number $z = x \cdot 10^k$ with $k \in \mathbb{Z}$ to be in the interval $[1, 10]$. Scale invariance for $p(x)$ then means, that the probability $p(x)\,\mathrm{d}x$ of $x \in [x, \mathrm{d}x]$ is equal to the probability $p(\lambda x)\,\mathrm{d}(\lambda x)$ of $\lambda x \in [\lambda x, \mathrm{d}(\lambda x)]$ for the scaled probability density $p(\lambda x)$. From the normalization condition

$$\begin{aligned}\int_1^{10} p(x)\,\mathrm{d}x &= \int_1^{10} p(\lambda x)\,\mathrm{d}(\lambda x)\\ &= \lambda \int_1^{10} p(\lambda x)\,\mathrm{d}x \overset{!}{=} 1\end{aligned} \tag{1}$$

of probability densities, the relation

$$p(\lambda x) = \lambda^{-1} p(x) \tag{2}$$

follows. Differentiation

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} p(\lambda x) = x \frac{\mathrm{d}p(x)}{\mathrm{d}x}, \qquad \frac{\mathrm{d}}{\mathrm{d}\lambda}[\lambda^{-1} p(x)] = -\frac{1}{\lambda^2} p(x) \tag{3}$$

of both sides of eq. (1) leads to the differential equation

$$x \frac{\mathrm{d}p(x)}{\mathrm{d}x} = -\frac{1}{\lambda^2} p(x), \tag{4}$$

which can be solved by separation of variables. Dividing the differential equation by x and $p(x)$ and furthermore integrating with respect to x leads to

$$\int \frac{1}{p(x)} \frac{\mathrm{d}p(x)}{\mathrm{d}x}\,\mathrm{d}x = \int \frac{1}{p(x)}\,\mathrm{d}p(x) = -\frac{1}{\lambda^2} \int \frac{1}{x}\,\mathrm{d}x. \tag{5}$$

Performing the indefinite integrals, one obtains

$$\ln[p(x)] = -\frac{1}{\lambda^2} \ln(x) + c, \quad c \in \mathbb{R}. \tag{6}$$

A this stage, one can set $\lambda = 1$ and raise both sides of the equation to the power of Euler's number, which yields

$$p(x) = \tilde{c} \frac{1}{x}, \tag{7}$$

where $\tilde{c} = e^c$. This constant $\tilde{c}$ is given by the normalization condition $\int_1^{10} p(x)\,\mathrm{d}x \overset{!}{=} 1$ as

$$\tilde{c} \int_1^{10} \frac{1}{x}\,\mathrm{d}x = \tilde{c} \ln(10) = 1 \quad \Leftrightarrow \quad \tilde{c} = \frac{1}{\ln(10)}, \tag{8}$$

hence the Benford probability density $p(x)$ is given by

$$p(x) = \frac{1}{\ln(10)} x^{-1}. \tag{9}$$

The probability $P(d)$ that the first digit of a number $z = x \cdot 10^k$ is the integer $d \in \{1, 2, \ldots, 9\}$ is therefore given by

$$\begin{aligned} P(d) &= \int_d^{d+1} p(x)\,\mathrm{d}x = \frac{1}{\ln(10)} \int_d^{d+1} \frac{1}{x}\,\mathrm{d}x \\ &= \frac{1}{\ln(10)}[\ln(d+1) - \ln(d)] \\ &= \frac{1}{\ln(10)} \ln\left(1 + \frac{1}{d}\right). \end{aligned} \tag{10}$$

Let a, b and c be numbers, such that $\log_a(c) = b$ is a well-defined expression. Then, $a^b = c$ holds and hence also $\log_b(a^b) = \log_b(c) = b \log_b(a)$ must hold. From this, the change of basis rule

$$\log_a(c) = \frac{\log_b(c)}{\log_b(a)} \tag{11}$$

follows. Using this rule, one obtains $\ln(10)^{-1} = \log(e)$ and $\ln(1 + 1/d) = \log(1 + 1/d)/\log(e)$. Thus, the final expression

$$P(d) = \log\left(1 + \frac{1}{d}\right) \tag{12}$$

for the Newcomb-Benford law is obtained. A dataset $Z \doteq \{z_1, \ldots, z_n\}$ consisting of $n \in \mathbb{N}$ numbers is said to follow a Benford distribution, if

$$\frac{n_d}{n} \approx P(d) = \log\left(1 + \frac{1}{d}\right) \tag{13}$$

for all $d \in \{1, 2, \ldots, 9\}$ holds, where n_d is the number of elements in Z with first digit $d \in \{1, 2, \ldots, 9\}$.

3 Methods

3.1 Generated datasets

This section describes two methods for generating datasets, which should follow a Benford distribution. These methods were implemented and examined as part of this work.

3.1.1 Method 1: Sampling from uniform distributions

A method directly derivational on the premiss of scale invariance is based on sampling from uniform distributions of various intervals. This method is illustrated with fig. 1 and is based on drawing samples

$$z_j \sim U(0, ka), \quad k \in \{1, \ldots, \lambda\} \tag{14}$$

with $j \in \{1, \ldots, n\}$ for $n \in \mathbb{N}$ from uniform distributions $U(0, ka)$. The probability that a sample z_j was drawn from the uniform distribution $U(0, ka)$ is herewith assumed to be

$$P(z_j \sim U[0, ka]) = \frac{1}{k} \tag{15}$$

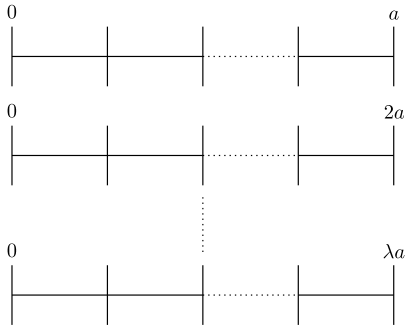

Figure 1: Dataset construction by sampling from uniform distributions of intervals $[0, a], [0, 2a], \ldots, [0, \lambda a]$.

to account for the overlap of the various uniform distributions. Given such a construction, one would expect that the resulting dataset should follow a Benford distribution, since it is constructed solely on the assumption of scale invariance.

3.1.2 Method 2: Sampling from a logarithmic uniform distribution

There is an instructive way to generate a dataset following a Benford distribution. First of all, one has to recall, to which type of data Benford's law applies. Benford's law applies to data spanning many orders of magnitude, since scale invariance was required for the derivation. Data covering many orders of magnitude is typically processed on a logarithmic scale as shown in fig. 2. Now, an a priori assumption on a dataset spanning many orders of magnitude would be, that it is uniformly distributed over those magnitudes. This is to say, that

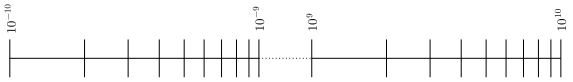

Figure 2: Logarithmic scale with base 10, ranging from 10^{-10} to 10^{10}.

such a dataset can be generated by sampling data from a range $[10^{r^-}, 10^{r^+}]$ on a logarithmic scale. A sample z_j for $j \in \{1, \ldots, n\}$ with $n \in \mathbb{N}$ therefore is obtained as

$$z_j = 10^{r_j}, \quad r_j \sim U(r^-, r^+), \tag{16}$$

where $U(r^-, r^+)$ denotes the uniform distribution between r^- and r^+ for r^-, $r^+ \in \mathbb{R}$. For an accordingly generated dataset, $\lim_{n\to\infty} \frac{n_d}{n} = P(d)$ should hold.

3.2 World population dataset

The current section describes the Newcomb-Benford law applied to investigate a real-world dataset, namely a dataset of country populations. The used country population dataset[1] features population data for all countries

[1]Data source: World Bank Open Data, `https://data.worldbank.org/indicator/SP.POP.TOTL`, last accessed: June 09, 2024.

on the world ranging from the period of 1970 to the year 2023. In a first step, only the first digits of country populations of one year (2022) were graphed and compared to the exact Newcomb-Benford law, whereas in a second step all data from 1970 to 2023 was used to compare it against the Benford law.

4 Results

4.1 Generated datasets

This section is concerned with presenting the results obtained for the generated datasets. Some key features and interesting properties are pointed out, which will be discussed in the section below.

4.1.1 Method 1: Sampling from uniform distributions

In fig. 3, the results obtained for method 1 of dataset generation are presented, namely those for the sampling from uniform distributions of various intervals. One can see, that for digits $d \in \{1, 2, 3\}$, the generated dataset seems to have a lower probability of occurrence than the Benford law would predict. However, for digits 4 to 9, the opposite is true; the predicted probabilities as predicted by the NBL are lower in this case than those present in the generated dataset.

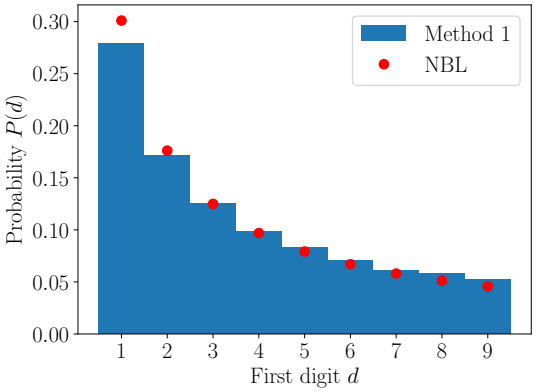

Figure 3: Comparison of the exact Newcomb-Benford law with a generated dataset according to method 1.

4.1.2 Method 2: Sampling from a logarithmic uniform distribution

The visualization fig. 4 features a histogram showing the distribution of first digits d in the generated dataset according to method 2; namely sampling from a logarithmic uniform distribution. As compared to method 1, this second method provides results even more coherent with the exact Newcomb-Benford law. For all first digits $d \in \{1, \ldots, 9\}$, the generated dataset seems to follow a Benford distribution very well.

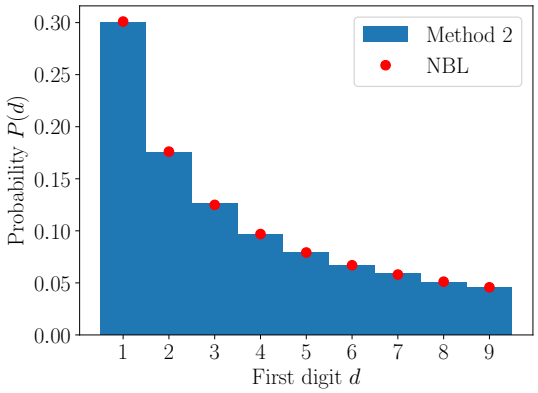

Figure 4: Comparison of the exact Newcomb-Benford law with a generated dataset according to method 2.

4.2 World population dataset

The world population data as obtained for only one year seems to only follow a Benford distribution roughly, as fig. 5 indicates. However, if data of all years present in the used dataset are taken into account, a different picture emerges: The world populations from years 1960 to 2023 seem follow a Benford distribution quite well, as fig. 6 shows.

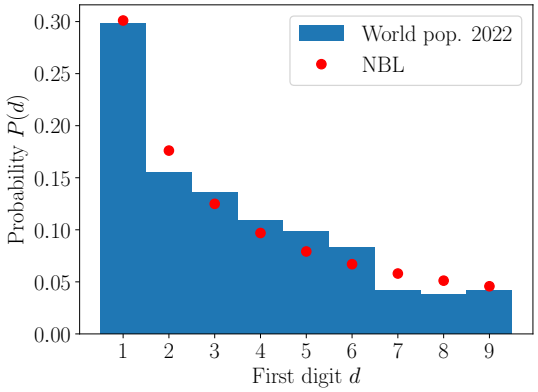

Figure 5: Comparison of the exact Newcomb-Benford law with the world population dataset for country populations of the year 2022.

5 Discussion

With respect to the two presented methods for generation of Benford distributed datasets one can say, that both methods seem to indeed produce datasets following a Benford distribution. Method 2 hereby provides better agreement with the exact NBL than method 1 does. Likely this could be explained by mathematically showing that the sampling method provides faster convergence to a Benford distributed dataset than method 1. One can propose the conjecture that in the limit $n \to \infty$ for n the amount of datapoints both methods will lead to exact Benford distribution; this presumption

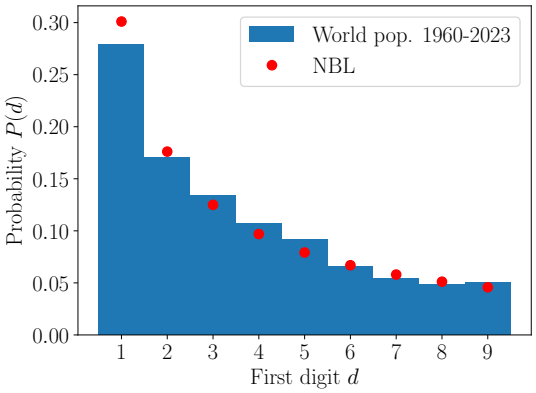

Figure 6: Comparison of the exact Newcomb-Benford law with the world population dataset for country populations from 1960 to 2023.

is substantiated by the presented results.

The application of the NBL to real-world data such as the world country populations provides insightful results. It is to be expected from the theory of Benford distributions, that a dataset complying with the requirements of a Benford distribution will approach an exact Benford distribution with increasing size of the dataset. As the world populations dataset matches the Benford distribution requirements, this is indeed what can be seen in the results. Benford's law indeed proves to be a powerful tool to check, if a given real-world dataset matching the Benford distribution requirements has a "natural" origin or not.

6 Conclusions

It was shown in this paper, that there are instructive ways to generate a Benford distributed dataset based only on the assumption of scale invariance. The generated test datasets according to both proposed methods showed a nearly Benford distributed behaviour.

In addition to this, the NBL was successfully applied to a dataset containing world country population numbers. Since this is a dataset spanning many orders of magnitude and because the dataset has natural processes, namely reproduction and migration, at its base, the set is expected to follow a Benford distribution. Indeed, this could be shown in the work at hand.

In conclusion, Benford's law provides an instructive and most interesting tool to detect, if a dataset spanning many orders of magnitude is of natural origin; that is to say, if natural processes have led to the resulting distribution of first digits in the dataset.

References

[Benford, 1938] Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572.

[Burgos and Santos, 2021] Burgos, A. and Santos, A. (2021). The Newcomb–Benford law: Scale invariance and a simple Markov process based on it. *American Journal of Physics*, 89(9):851–861.

[Newcomb, 1881] Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, 4(1):39–40.