The Newcomb-Benford Law, Applied to Binary Data: An Empirical and Theoretic Analysis

Gabriel Chuang (gtchuang@andrew.cmu.edu), Brandon Wu (bjwu@andrew.cmu.edu)
Carnegie Mellon University

***Abstract*— The Newcomb-Benford Law, which states that natural datasets tend to have many values with first digit 1, is often used for verification and auditing of potentially-fraudulent data. Given that much modern data is stored in binary, it important that this principle be applicable to binary data. We propose an extension, the Strong Newcomb-Benford Law, present several examples on real datasets, and discuss some implications.**

I. Introduction

Benford's law, sometimes called the Newcomb-Benford law, is an empirical observation about many sets of real-life numerical data. It was discovered by Simon Newcomb in 1881, and then rediscovered by Frank Benford in 1938.

In the spirit of fairness, we will alternate between calling it Benford's Law and Newcomb's law.

In the digital age, almost all data is represented in binary. In this paper, we discuss the application of Newcomb's law to data represented in binary form.

II. Background: Benford's Law

Newcomb's law notes that the distribution of leading digits in naturally-occurring data is skewed towards smaller digits. For example, consider the populations of the world's countries. Of the 237 countries[1], 64 (approximately 27%) have a population with 1 as the first digit. In contrast, only 12 have a population where 9 is the first digit.

Benford's law predicts that 1 is the most significant digit in approximately 30% of datapoints, while 9 is the most significant digit in approximately 5% of datapoints. Newcomb's law is observed to occur regardless of scale (e.g. changing the units on the data still results in the same distribution).

This results in a distribution like that displayed in Fig. 1.

Benford's law has been observed to hold on a multitude of datasets, such as:

- Country populations
- Building heights
- Molecular weights
- Election data and vote counts
- Daily volume of diet coke consumed by John Mackey
- Powers of 2
- Fibonacci numbers

Notably, it does *not* occur in some other distributions, especially ones where the data is not approximately exponentially distributed, such as:

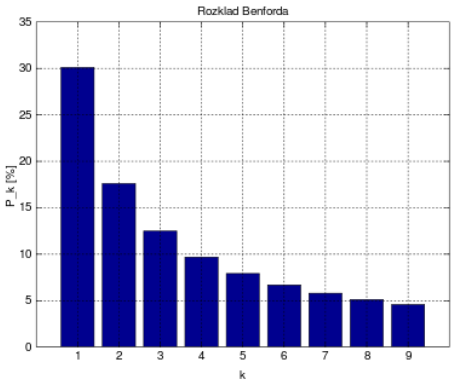

Fig. 1: Distribution of first digits found in many datasets, per Benford's law. Note that larger digits rarely appear as the first digit in many datasets. Source: Wikipedia.

- Height and weight
- Amount of wood that a woodchuck would chuck if a woodchuck could chuck wood
- Scores on 15-210 exams

There are many explanations for why Newcomb's law occurs, most of them centering on logarithms and how exponentially-distributed data tends to favor leading-digit 1's. In fact, some mathematical datasets like the powers of 2 are proven to converge on the ratios suggested by Benford's law. For brevity, those explanations are omitted here. Insightful discussions can be found here.

Benford's law has found extensive use in fraud and forgery detection, since inexperienced data-fakers may present data that is not distributed as suggested by Newcomb's law.

III. Motivation

To date, Benford's law has only been extensively discussed on datasets represented in decimal (base-10). Nowadays, however, binary (base 2) has supplanted decimal as the favored method of storing data efficiently, with nearly all of society relying in some form on binary data[2].

With the rise of increased reliance of technology, however, Bad People like The Russians and other Scary CyberCriminals have become a major threat, in part because they can inject false data into our datasets. To guard against this, it is imperative that we investigate the applicability of Newcomb's law to binary data, to allow us to extend the use

[1]Technically, "countries and dependencies", as determined by the international nongovernmental organization known as Wikipedia.

[2]With the notable exceptions of grade school children, engineers, and the Sentinelese.

of Benford's law to guard against digital data forgery and fraud.

IV. The Strong Newcomb-Benford Law

We claim that a stronger version of Benford's law holds on binary data. Specifically, we claim that ***All* datasets with nonzero elements have a 100% incidence of 1 occurring as the first digit.**

Note that this is stronger than Newcomb's law in base 10, which only claims approximately a 30% incidence of 1 occurring as the first digit.

V. Theory

To substantiate our claim, we will now prove the following theorem.

Theorem 1: All non-zero numbers have a leading one in their base-2 representation.

Unfortunately, due to a lack of appreciation for our work on the part of NSF grant committees, we lack the funds to formally prove the theorem. We will instead have a Concepts student prove the theorem. The proof is below. Apologies in advance.

Proof: Let n be a number in the set N. By the contrapositive, either this number is zero or it isn't. If it's zero, then WLOG the theorem is true. Otherwise, we need to show that the first digit in the binary representation is 1. First, let's case on where the number is. If it's between 1 and 2, then we know the binary representation is either 1 or 10, and both of those start with 1. Otherwise, inductively, it's more than 2, then, if it's between 2 and 4, not including 2, it must be 2 or 3 digits long. The first digit can't be 0, because otherwise it would be $1 or 2 digits long$. We can easily see that this would be true for any length, not just 2 or 3, so we have successfully proved what we want to show. $\implies \impliedby$ ■

VI. Empirical analysis

If you're like us[3], you probably weren't convinced by the above proof. But, as we all know, numbers don't lie. We collected several exhaustive sets of data from reliable sources and plotted the proportion of data points with first digits 0 and 1 respectively. The sources are listed alongside the figures.

The tabulated data is omitted, for brevity, as well as to protect the privacy of our data sources. We converted all of the data points to their binary representations, and plotted the relative frequency of the leading digits. The figures are shown in Figs. 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, and 14 in the appendix[4].

VII. Conclusions

In conclusion, as supported by both a mathematically rigorous proof and large quantities of collected empirical data, we verify that the Strong Newcomb-Benford law holds on nearly all datasets, except on datasets with high proportions of zero values. This suggests a few applications for the area of detecting various Bad Things, such as fraud or malicious injection attacks. Specifically, we can check the veracity of datasets by checking that all numbers in those datasets begin with 1 when represented in binary.

VIII. References

Our references have been redacted to protect the identities of our data sources.

IX. Acknowledgements

We are very thankful to [redacted] for providing us with the entire codebases of several large tech companies. We also thank [redacted] for measuring the height of every chair in Gates for us for two slices of pizza and a soda[5]. Last but not least, we would like to thank the FBI for sharing their secretly-recorded audio files of crying students with us. Without their generously-provided data, this project would not have been possible.

X. Appendix: Figures

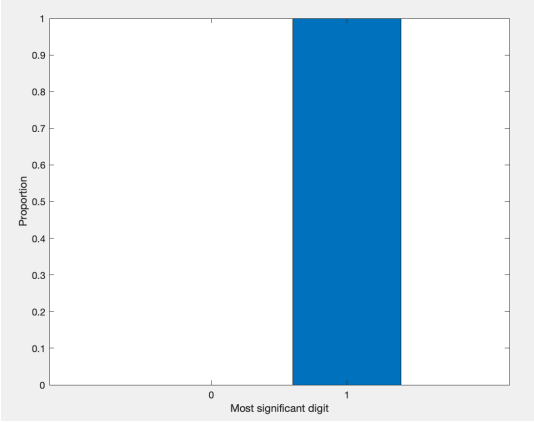

Fig. 2: Frequency of leading digit in populations of 237 nations and dependencies, as listed by Wikipedia, when represented in binary. Source: Wikipedia

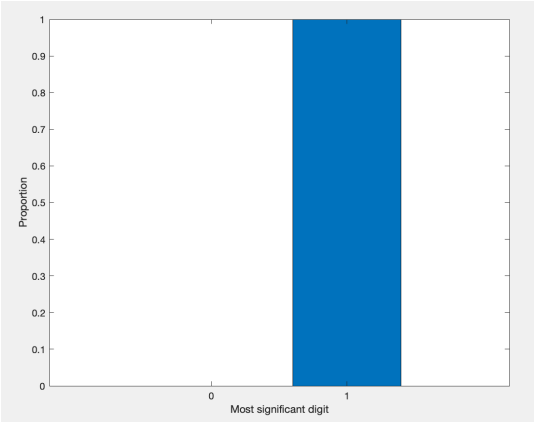

Fig. 3: Frequency of leading digit in molecular weight of 1800 chemicals, when represented in binary. Source: "The Law of Anomalous Numbers", F. Benford.

[3] A frightening thought.

[4] You can tell we're legit by how many figures our paper has.

[5] Starving upperclassmen will do anything for "free" food.

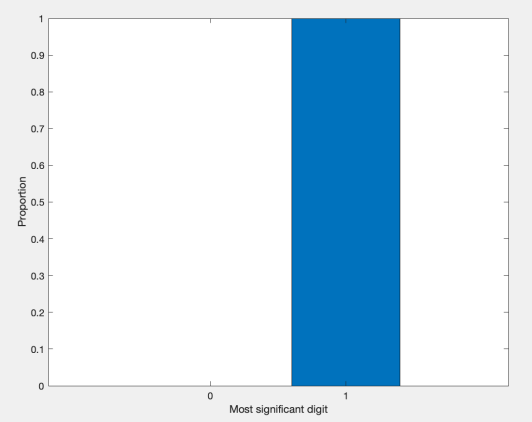

Fig. 4: Number of sunny days in Philadelphia, per year, 1735-2020, when represented in binary. Source: weather.com

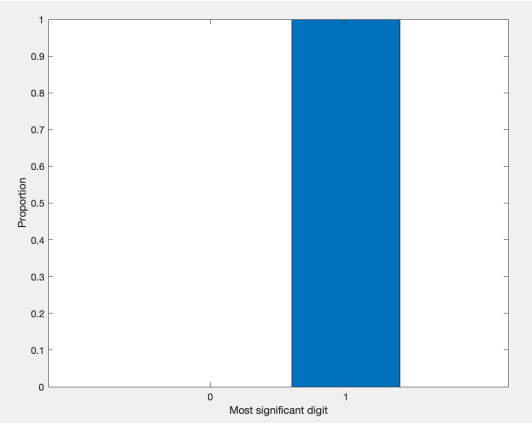

Fig. 5: Frequency of leading digit in height, in millimeters, of chairs in CMU's Gates-Hillman Center, when represented in binary. Source: original research

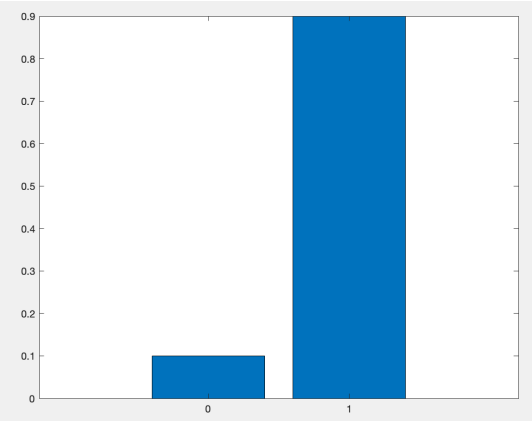

Fig. 6: Frequency of leading digit in binary representation of MNIST handwritten digit datasets. Source: MNIST

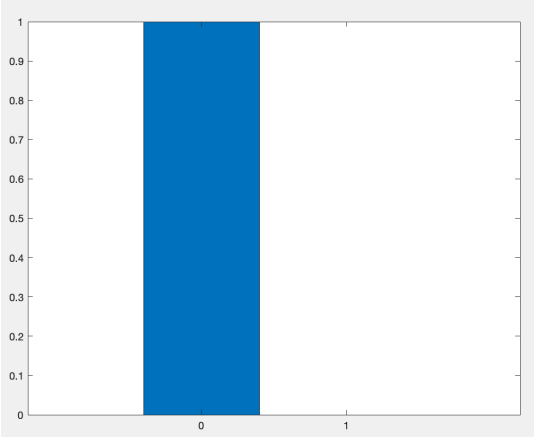

Fig. 7: Frequency of leading digit in number of lines of code written in Standard ML in tech industry codebases, when represented in binary. Source: anonymous

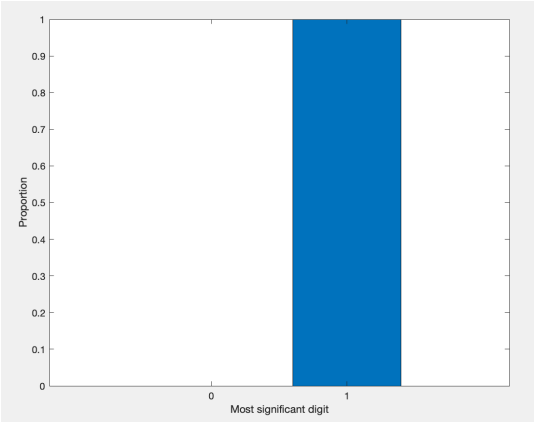

Fig. 8: Frequency of leading digit in $GME stock share price over the first quarter of 2021, when represented in binary. Source: NYSE

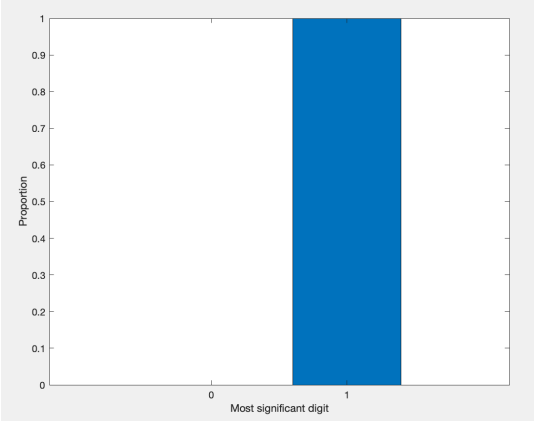

Fig. 9: Frequency of leading digit in 15-210 exam scores, when represented in binary. Source: Anonymous.

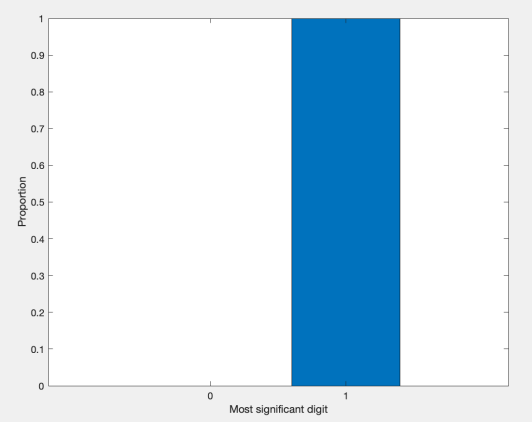

Fig. 10: Frequency of leading digit in number of Asian-American children with first name Kevin enrolled in public high schools in the Bay Area during the 2012-13 school year, aggregated at a county level, when represented in binary. Source: [redacted], Office of the California Department of Education

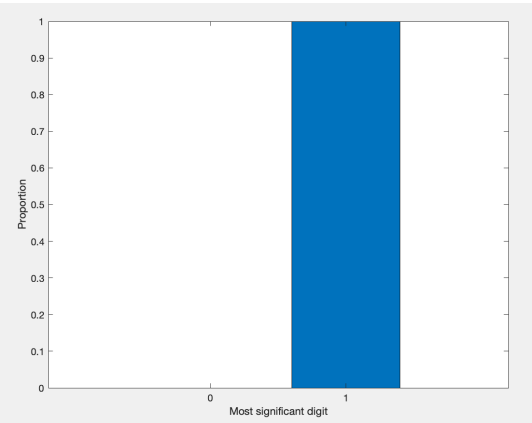

Fig. 11: Frequency of leading digit of volume, in decibels, of 435 audio samples of students crying after calculus exams, when represented in binary. Source: FBI

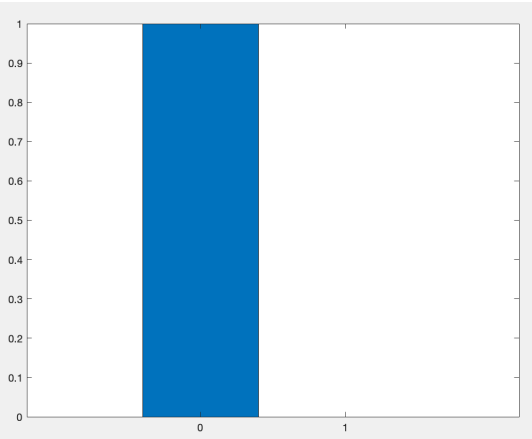

Fig. 12: Frequency of leading digit of student support for tents on the College of Fine Arts lawn, in percent, when represented in binary. Source: CMU S3

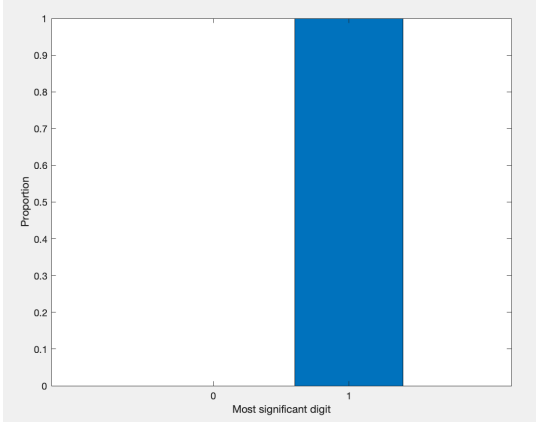

Fig. 13: Frequency of leading digit of compiled binaries of several viruses, malware, and worms available on [redacted]. Source: [redacted]

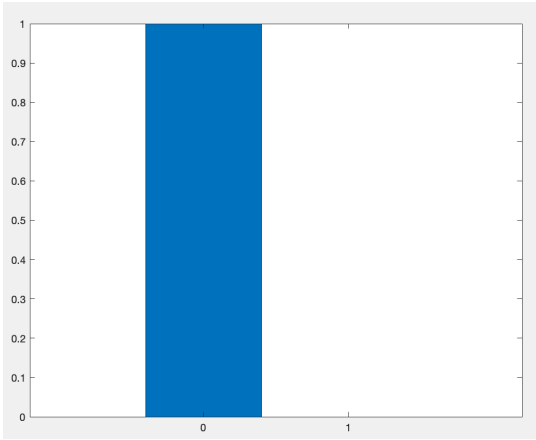

Fig. 14: Frequency of leading digit in daily counts of number of students with cameras on during Zoom classes. Source: Original research.