Benford's Law and Mathematics as a Political Weapon

Jackson Howell

May 4, 2022

Abstract

This paper explains the history of Benford's Law, a statistical observation concerning the leading digits of many real-world datasets. We briefly describe its technical definition and describe how the uniform distribution of a random variable in logarithmic space provides a sufficient condition for fulfilling Benford's Law. We give examples of real-world applications of Benford's Law and how these applications may be problematic. In particular, we identify the use of Benford's Law to prove election fraud as particularly fraught. We conclude by relating discussion from other students on the topic of the politicization of mathematics.

1 Introduction

In 1881, Canadian-American astronomer Simon Newcomb published a paper in which he remarked that the first pages of logarithm books (especially those associated with the logarithms of numbers beginning with 1) were more worn out than the later ones (Newcomb [1881]). This publication was one of the first recorded observations of what was later formalized by American physicist Frank Benford in 1938 (Benford [1938]). His finding, which has since become known as Benford's Law, notes that in many real-world datasets, the first digit of any observation is more likely to be 1 than 2, more likely to be 2 than 3, and so on. He even specified exact percentages for each leading digit. He observed conformity to these proportions in data sources as disparate as the areas of rivers, the populations of American settlements, atomic weights, and the numbers printed in a year's worth of *Reader's Digest*.

Benford's Law has become a useful tool outside of mathematics. Forensic accountants often use it as a first-sweep tool to detect suspicious activity. However, it has been widely and incorrectly applied to assert fraud in American elections. How can Benford's Law, or any other scientific phenomenon, be explained to a popular audience with little mathematics education? Is mathematics doomed to becoming just another front in ever-growing political conflicts? In this paper, I present the mathematical explanation of Benford's Law, positive and negative applications of the law, and a discussion of mathematics in politics.

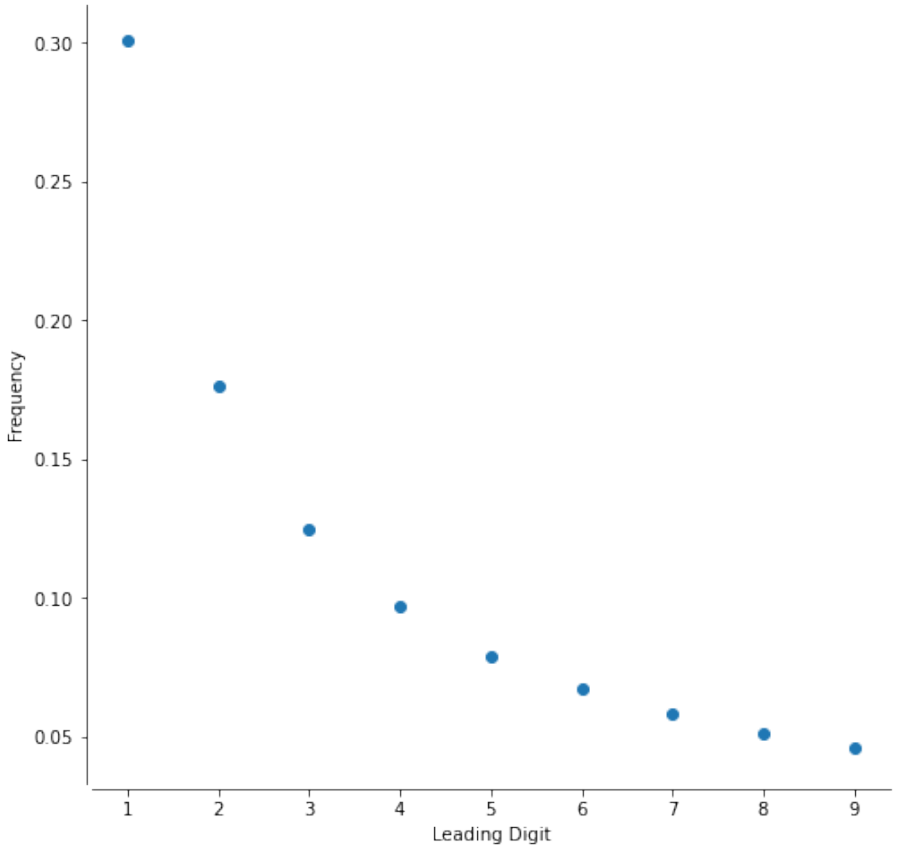
2 The Mathematics of Benford's Law

2.1 Definition

Benford's Law is fulfilled for a dataset when the leading digits of the data are distributed according to the function

$$P(d) = \log(d+1) - \log(d),$$

Figure 1: Frequency of Leading Digits under Benford's Law



Leading Digit	1	2	3	4	5	6	7	8	9
Frequency	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

where $d \in \{1, \ldots, 9\}$ and log denotes $\log_{10}$, which will be true throughout this paper. This function is graphed in Figure 1. The exact values are also provided below that figure.

Since these probabilities are positive and sum to 1, this function is actually a discrete probability measure on the support $\{1, \ldots, 9\}$. Furthermore, the Benford Probability Mass Function can be interpreted as a statement about the distribution of leading digits in logarithmic space (Fewster [2009]). The expression $\log(d+1) - \log(d)$ can be read as a distance, such that Benford's Law is satisfied when each digit occurs with probability equal to the distance between it and it's successor on a logarithmic scale. Therefore, Benford's Law is satisfied exactly when leading digits are distributed uniformly on a logarithmic scale.

2.2 Example: the Reciprocal Distribution

Consider the reciprocal distribution, which is defined by the PDF

$$f(x; a, b) = \frac{1}{x \log \frac{b}{a}}, \qquad x \in [a, b], \qquad 0 < a < b.$$

This distribution is alternatively known as the log-uniform distribution, since for any X distributed according to the reciprocal distribution, $\log X$ is distributed uniformly. We generate a 10,000-size vector of i.i.d. random variables distributed log-uniform from 0.01 to 1000. The distribution of the resulting vector is given in log-linear and linear-linear scales in Figure 2.

The random vector is indeed distributed uniformly on a logarithmic scale. On a linear scale, we see that values closer to 0 are far more likely than greater values. How are the leading digits distributed? In Figure 3, the frequencies of leading digits are given along with the expected values from Benford's Law, and we can see that these data conform very closely to those expectations.

Why do log-uniform variables seem to comply closely with Benford's Law? A log-uniform random variable X with support from 0.01 to 1000 takes on values within a given range with probability

$$\frac{1}{\log 1000 - \log 0.01} \left(\log(b) - \log(a)\right) = \frac{1}{5}(\log(b) - \log(a)).$$

The variable X takes on a leading digit of 1 when $0.01 \leq X < 0.02$, or $0.1 \leq X < 0.2$, or $1 \leq X < 2$, or $10 \leq X < 20$, or $100 \leq X < 200$. Each of these probabilities are therefore $\frac{1}{5}(\log(2 * 10^k) - \log(1 * 10^k)) = \frac{1}{5}(\log 2 - \log 1)$. Therefore, the probability that X takes on a leading digit of 1 is $5\frac{1}{5}(\log 2 - \log 1) = \log 2 - \log 1 \approx 0.301$, which is exactly the Benford prediction. This pattern holds true for the other digits as well.

This is not intended as a formal proof, but rather as an intuitive one: when a random variable is distributed uniformly on a log-scale, so are its leading digits. Therefore, random variables that are close to a broad, uniform distribution in logarithmic space are more likely

Figure 2: Distribution of Log-Uniform Random Vector X

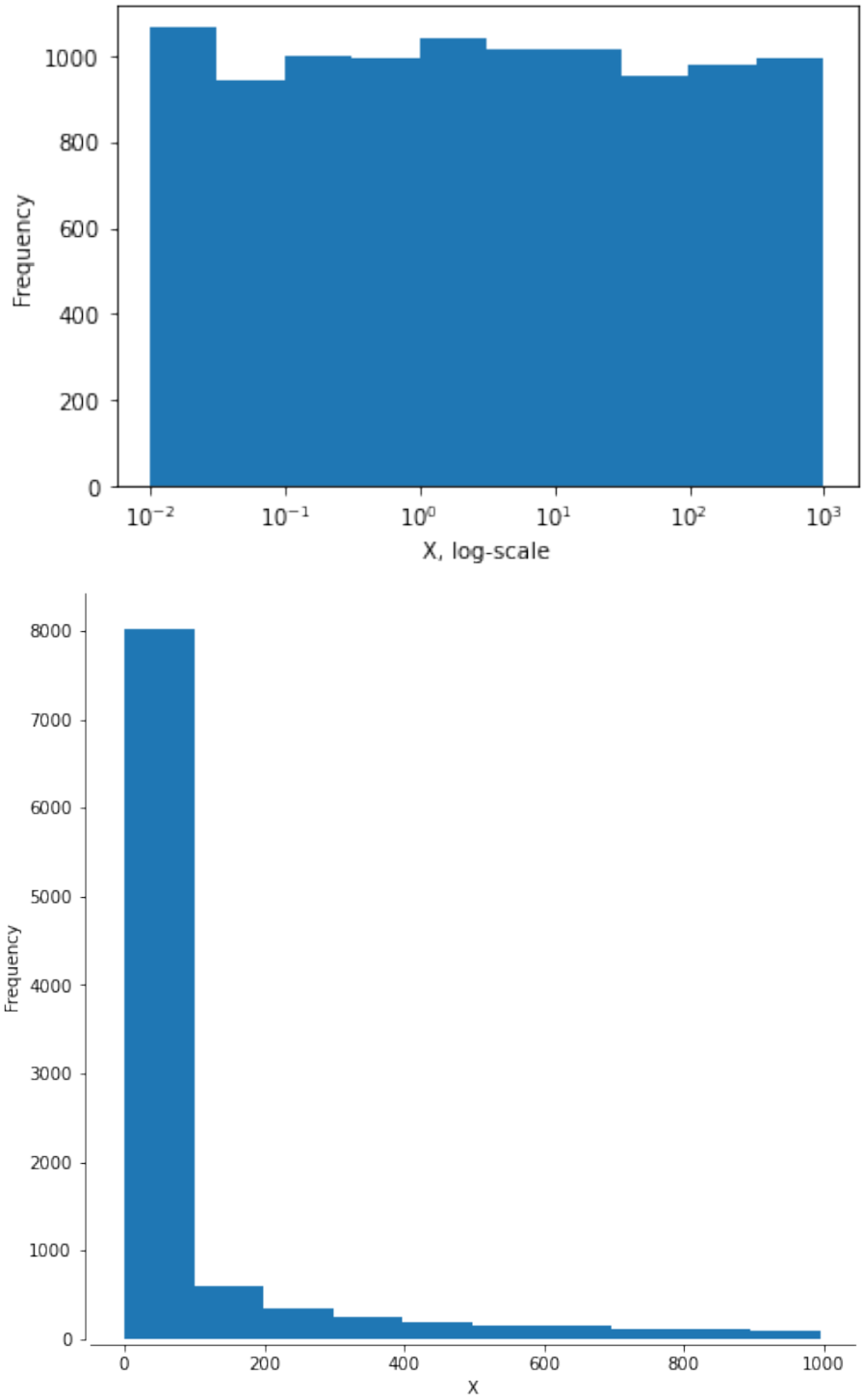
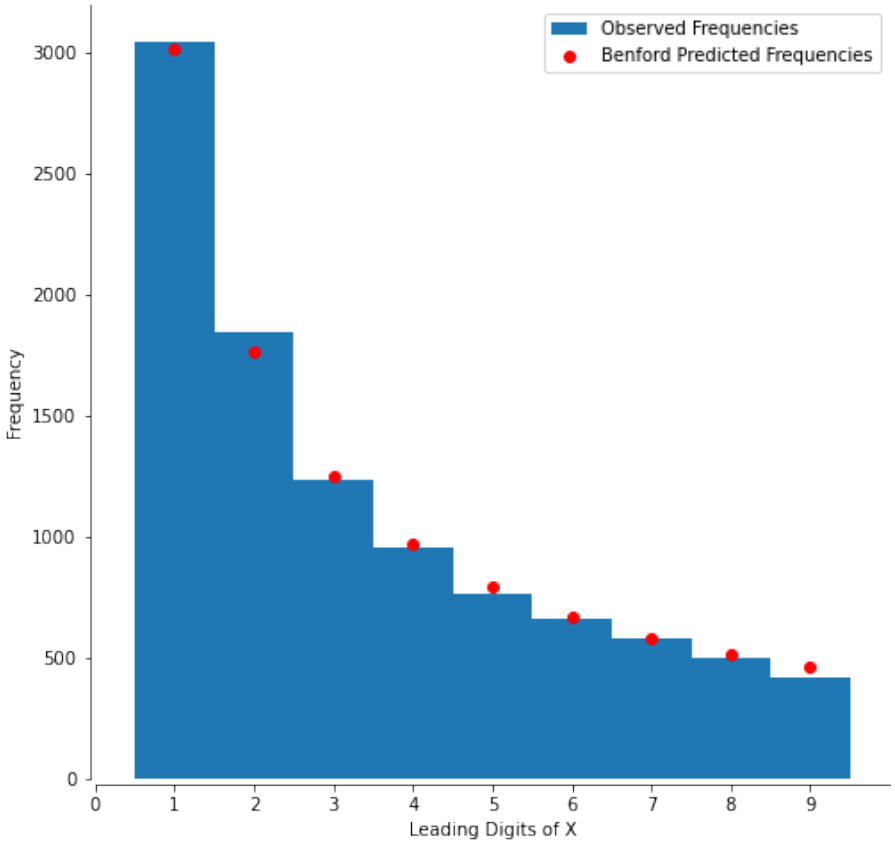


Figure 3: Leading Digits of Log-Uniform Random Vector X



to follow Benford's Law. The location of the endpoints, the closeness to uniform distribution, and the number of sample sizes will affect divergence from Benford's Law.

Therefore, we should not expect all data to conform to Benford's Law: distributions close to the reciprocal distribution (such as the exponential distribution) will conform well, while other distributions (such as the normal) will not. For instance, Figure 4 shows the distribution of a 10,000-size vector of i.i.d. Normal(1000, 100) variables in logarithmic space, and the distribution of leading digits. We can observe the narrow distribution of these data in logarithmic space and complete lack of conformity to Benford's Law.

3 Applications

3.1 Forensic Accounting

Benford's Law has become increasingly popular as a tool for forensic accountants to detect fraud during audits. Under the assumption that accounting data should follow Benford's Law, by finding the proportions of leading digits, accountants can conduct a first sweep to find any suspicious patterns. Significant divergences from Benford's Law might raise questions as to whether data are being artificially manipulated.

For instance, accountant Pete Miller (Miller [2016]) describes how he has used Benford's Law in an investigation. Providing the graph in Figure 5, he notes the statistically significant divergence from Benford's Law exhibited in digits 1 and 3. Focusing on the specific entries that began with those digits, he was able to find suspicious checks made out to a specific vendor, which led to a more detailed investigation.

Pete Miller implicitly invoked the assumption that whatever data he was reviewing should comply to Benford's Law. Accounting data often does conform to a reciprocal-like distribution: most expenditures or revenues are small, while fewer are very large. This pattern is more likely to conform to a log-uniform distribution. Even with this assumption, divergence from Benford's Law was not airtight evidence for wrongdoing. It was merely a starting point

Figure 4: Distribution of Normal Random Vector X and Leading Digits

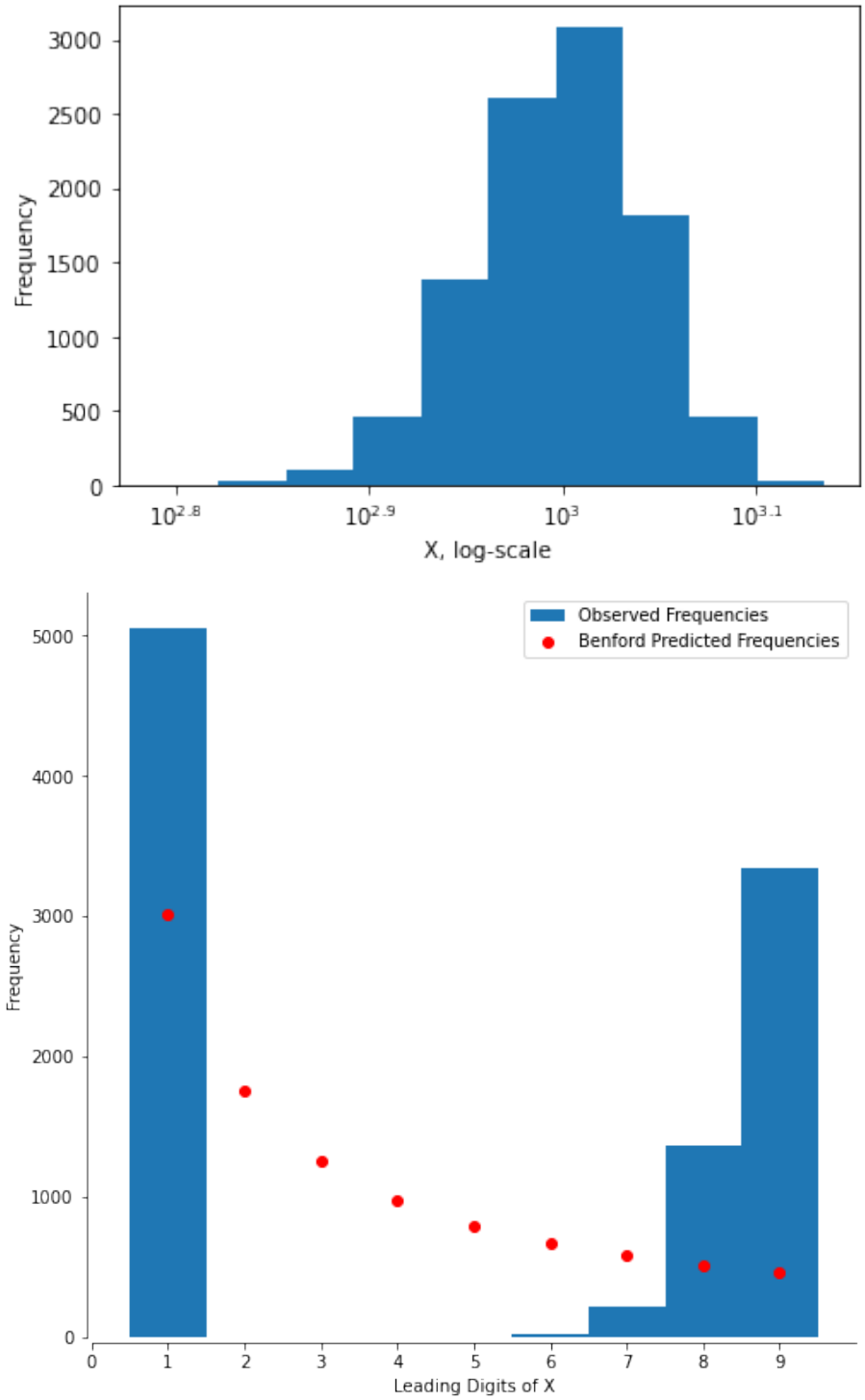
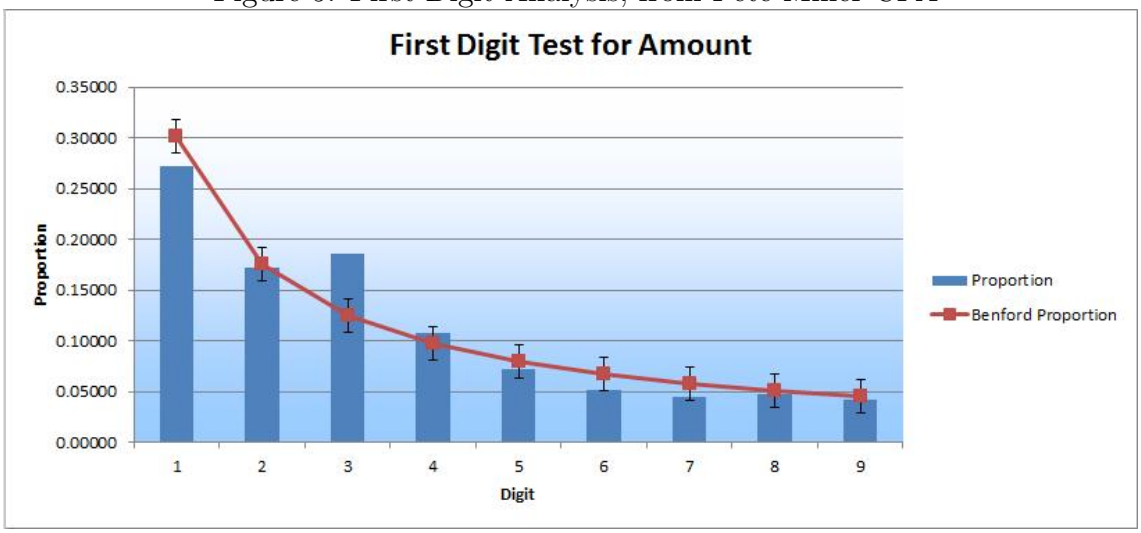


Figure 5: First-Digit Analysis, from Pete Miller CPA



that led to further investigation, which may or may not have uncovered fraud.

3.2 2020 Election Returns

Benford's Law was invoked as evidence of fraud during the 2020 United States Presidential Election. I use data and figures provided through an analysis of this online debate by Walter Mebane, Jr., a Professor of Political Science and Statistics at the University of Michigan (Walter R. Mebane [November 2020]).

Figure 6 shows the leading digits of precinct returns from the Chicago area between the Biden/Harris and Trump/Pence tickets. Figures similar to this were widely published online as evidence of fraud on the part of the Biden campaign, since Trump's leading digits follow Benford's Law more closely than Biden's.

However, is there any reason to believe that these precinct results are distributed log-uniform and therefore that they should follow Benford's Law? By looking at the raw distributions of precinct returns in Figure 7, we see that Biden's votes are distributed approximately normally, while Trump's are distributed somewhat reciprocally. Therefore, the Republican vote totals, but not the Democrat totals, could be expected to follow Benford's Law.

Why were Trump's votes distributed this way and not Biden's? Chicago, as a Democrat

Figure 6: Chicago Precinct Returns, Leading Digits

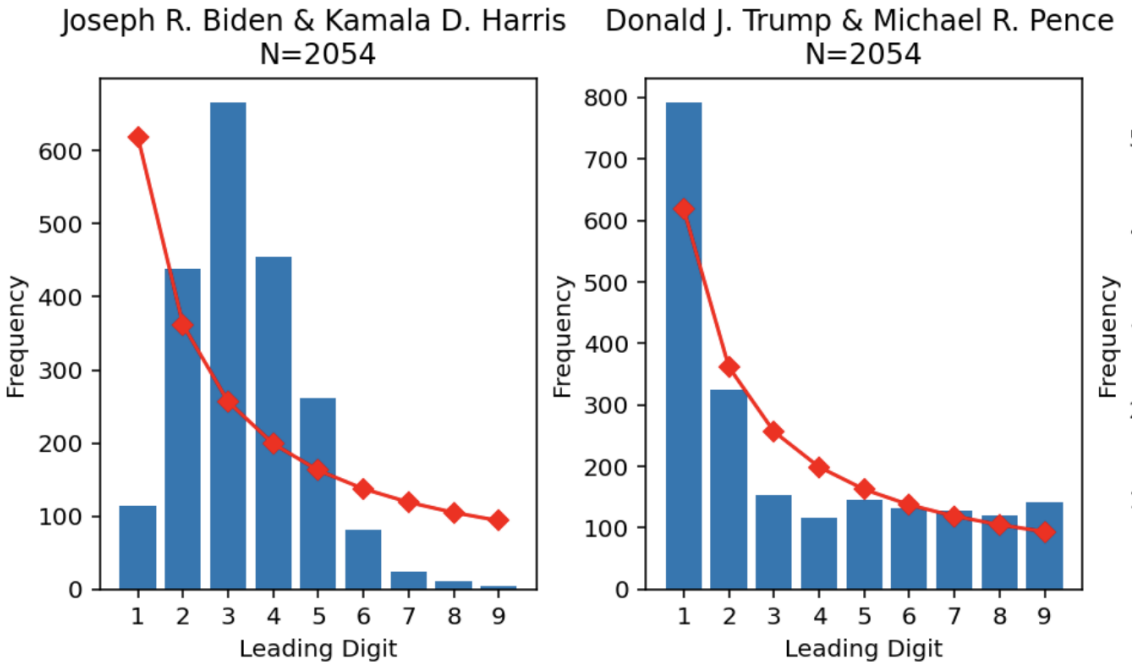
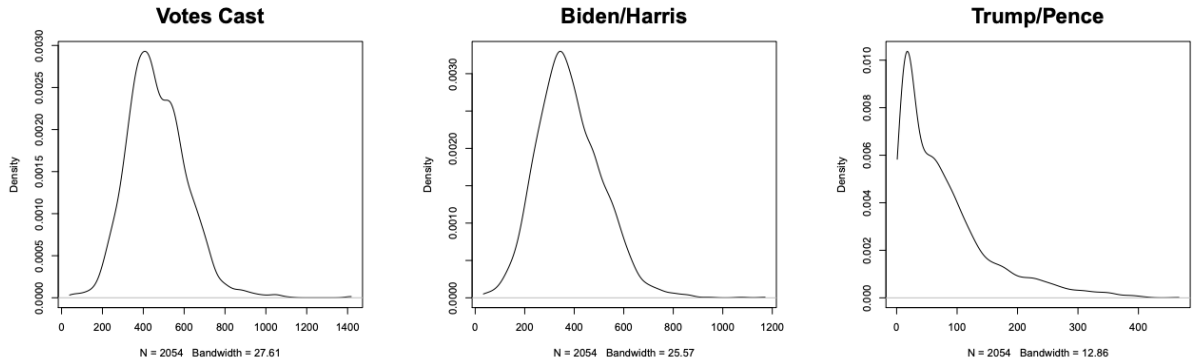


Figure 7: Chicago Vote Returns, Total



stronghold, is very pro-Biden and anti-Trump. According to Mebane, Biden/Harris received an average of 82% of the vote in these precincts, while Trump/Pence received 17%. Both of these distributions were skewed: with many districts reporting near 100% for Biden/Harris and 0% for Trump/Pence. Therefore, the distribution of votes is mostly a function of the distribution of precinct sizes.

Most precincts are meant to be the same size (around 500 voters or so), but they vary approximately normally. Since Biden/Harris won virtually all the votes in these precincts, his totals are therefore distributed mostly normally. On the contrary, the Trump/Pence ticket's poor performance in many precincts led to a high probability-density for vote returns near zero. These differences in distribution are therefore due to differences in popularity between the candidates in the Chicago area. There is no evidence for fraud from Benford's Law, therefore, since Biden's totals are not, and are not be expected to be, near-reciprocally distributed.

Benford's Law should never have been used to analyze these results. Indeed, it is well-known that Benford's Law is not applicable to analyzing the first digits of vote returns, since these are often not distributed log-uniform (Pericchi and Torres [2011]). Nevertheless, Benford's Law continues to be invoked as evidence for fraud. Recently, the loser of the 2021 California Governor Recall Election claimed fraud in results using Benford's Law, before any voting data had even become available (Mathis-Lilley [2021]). Clearly, it is becoming popular to invoke the law to prove election fraud regardless of whether it is well-applied to the data, and even regardless of whether the data exist yet!

4 Politicization and Discussion

In my presentation, I stressed the use of mathematics as a political weapon as a worrying development in public discourse. Math plays a unique, dual role in the public imagination. First of all, mathematics is seen as profoundly true. Therefore, mathematical evidence is

extremely strong, because math is always logical. When George Orwell wanted to create a government so absurdly authoritarian that it could make its citizens believe in anything it wanted, he had the government proclaim 2+2=5. That the citizens accepted this was a signal to the reader of how enslaved the citizens' logical faculties were.

Mathematics is also a subject that most average people think is very confusing. Indeed, many proclaim being bad at math as something to be proud of. How can we expect citizens to be able to weigh mathematics as evidence if they do not understand any math beyond basic arithmetic and algebra? These points sparked much discussion on Blackboard, since our class has identified math education as an interesting theme throughout the semester. I present others' comments in italics, and my own comments, on these themes.

I think that a different strategy can be effective, maybe connecting the law to something people already know...Basically everybody has at least heard of the bell curve, and so pointing out that the law is so non-universal that it doesn't even apply there, I think, would...satisfy a lot of people.

This point identifies a useful strategy in education: don't tell people more than they want to know. While those with collegiate math experience might appreciate the discussion of the log-uniform distribution and of the distance interpretation of Benford's Law, most people would become lost. But, as the commenter points out, most people know about the bell curve, and by showing the first digits of some real-world normal dataset, Benford's Law can be made more concrete.

Indeed, by showing examples of how Benford's Law applies or does not apply to non-contentious datasets, this point could be made even more strongly. For example, while lengths of rivers in California would follow Benford's Law, the populations of that state's electoral districts would not. Because there should be no contention that these figures have been somehow meddled with, a simple figure showing how Benford's Law describes the former better than the latter would go a long way to disabusing people of the notion that every

dataset should be expected to conform to the law. This could also prevent the idea that we are simply trying to "cover something up" on narrow political grounds, by showing that there are non-political datasets that don't follow Benford's Law.

When you mentioned how people would use the law to make claims in media that weren't entirely true, my first thought was about vaccine / testing efficacy and how those stats can be misleading without proper analysis (mainly with conditional probabilities, like a 90% chance of a test being right if you actually have (COVID or whatever it may be)), stuff like that; Just another example of math used in the media that likely goes over most people's heads and gets misinterpreted.

Conditional probability is a nuanced thing even for people with some collegiate statistics education! These probabilities are often used to justify (or to justify ending) lockdowns, mask mandates, vaccine mandates, universal testing, etc. This also connects to the discussion on conditional probability and the Monty Hall Problem that another presentation touched on in the context of gambling.

Commenter One: *It reminds me of what my grandparents would send me: nonsense or at least poor-established "science research" that telling us what to do and what to avoid. Especially in such a fast-developing era, my question is then whether science can truly remain its neutrality not only in definition but also in reality. Swiping out science ignorance would always be a step behind the scientific development. This does not imply that science education is eventually pointless but is a great fundamental concern I have in mind.*

Commenter Two: *For me, the proposition of communicating or discussing a topic such as Benford's Law or climate change in a meaningful way with tens of millions of people is fantastical. It takes many, many years for a consensus on topics such as these to disseminate over any relevant subset of such a large population even when the science is incredibly consistent, and the impact of this dilemma will be felt more acutely as/if the internal political*

situation of this country becomes more volatile.

These two comments are rather pessimistic, but they seem to reflect the consensus opinion reached in the discussion board. For many, political discourse seems hopelessly toxic. We see falsified or misconstrued evidence used on many political fronts to justify all sorts of policies. All it takes is one Facebook meme to invoke Benford's Law, and now someone believes that there is mathematical evidence of fraud in the election. It is unlikely that a meme explaining the distributions underpinning Benford's Law would do as well. The problem gets even harder for ideas that are more complicated and that require more underlying knowledge.

5 Conclusion

It is unlikely that mathematicians alone can restore reason to modern political discourse. And, like it or not, mathematics has become a standard tool used and misused by political actors. Although many of us may feel dejected and hopeless as to how to rectify this situation, we should keep in mind a few positive observations.

First, mathematics retains its reputation as a supremely logical subject. Otherwise, nobody would feel the need to appeal to math in their arguments. Unlike science, which has come under attack for its findings on evolution and climate change, math remains (mostly) unscathed. As such, those of us with a mathematics education may still hold some authority and credibility when we correctly employ math to make a point.

Second, even though most people don't like mathematics, they still have some knowledge of it through school and popular culture. As the first commenter said, most people know about the bell curve. Because Benford's Law has a simple relationship with bell curves–i.e., it does not apply to bell curves–this finding could be rather easily taught to the wider population. As our class's many discussions on gambling have shown, most people have a basic knowledge of probabilities through things like card games. Mathematicians can use these cultural entry-points as footholds to pick apart faulty mathematical arguments.

Finally, it is important to remember that internet culture is not everything. Most people trust their friends and family more than some stranger on Facebook. One commenter spoke about their frustration with getting faulty scientific arguments from their grandparents. Although such messages may be infuriating, we should be grateful that our family/friends reach out to us, so that we maintain communication and mutual trust. By being a trusted, mathematically educated member of a close relationship, we can project positive and accurate mathematical thinking to those nearest to us. That is at least a start.

References

Frank Benford. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572, 1938. ISSN 0003049X. URL http://www.jstor.org/stable/984802.

R. M. Fewster. A simple explanation of benford's law. *The American Statistician*, pages 26–32, 2009. URL https://www.stat.auckland.ac.nz/~fewster/RFewster_Benford.pdf.

Ben Mathis-Lilley. Larry elder announced he "detected fraud" in california recall vote results when they didn't yet exist. September 2021. URL https://bit.ly/36tkRIu.

Pete Miller. Benford's law: A real life case study. January 2016. URL https://www.acfeinsights.com/acfe-insights/2016/1/15/benfords-law-a-real-life-case-study.

Simon Newcomb. Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, 4(1):39–40, 1881. ISSN 00029327, 10806377. URL http://www.jstor.org/stable/2369148.

Luis Pericchi and David Torres. Quick Anomaly Detection by the Newcomb–Benford Law, with Applications to Electoral Processes Data from the USA, Puerto Rico and Venezuela. *Statistical Science*, 26(4):502 – 516, 2011. doi: 10.1214/09-STS296. URL https://doi.org/10.1214/09-STS296.

Jr. Walter R. Mebane. Inappropriate applications of benford's law regularities to some data from the 2020 presidential election in the united states. November 2020. URL http://www-personal.umich.edu/~wmebane/inapB.pdf.