Testing Benford's Law With Data From The Mathematical World

Nolan Kriener
Faculty Mentor: Alex Smith

University of Wisconsin-Eau Claire Department of Mathematics
April 2013

1. Benford's Law

Benford's Law, also known as the First Digit Law, is a principle regarding a pattern occurring in large data sets. It states that if you were to look at a large set of numbers from the physical world (lengths of rivers, heights of trees, stock prices, etc.), and observe the first digit of each term in the set, a pattern emerges. Many might assume that the digits 1-9 would occur with roughly equal frequency, however this is not the case. In fact, the digits occur with varying probabilities, as given by[3]

$$P(d) = \log_{10}\left(1 + \frac{1}{d}\right)$$

where d is your digit in question, resulting in the following probability distribution:

$$\begin{pmatrix} \text{Digit} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ \text{Probability} & 30.1\% & 17.6\% & 12.5\% & 9.7\% & 7.9\% & 6.7\% & 5.8\% & 5.1\% & 4.6\% \end{pmatrix}$$

The aforementioned equation pertains to base 10, however Benford's Law applies to any base, where the only modification to the equation that is required is changing the base of the log to your base in question rather than 10. No matter what base is used, the resulting probability distribution follows a similar pattern, with 1 occurring most frequently, with each subsequent digit occurring less frequently, up until your final digit of $b-1$, which occurs least often.

2. The Physical World vs Mathematical World

One important aspect of Benford's Law is that it is exclusive to data sets from the physical world. Our research examines the prevalence of Benford's Law in data sets not from the physical world but rather from the mathematical world. To do this, we focused primarily on non-monotonic sequences. We could not focus on strictly increasing or strictly decreasing sequences as our end results would be skewed based on how many terms of the sequence we decided to look at. For this project, we found these non-monotonic sequences from a few sources, most commonly from the Online Encyclopedia of Integer Sequences, or OEIS for short[1]. The OEIS is also known as the Sloane Integer Sequence Database after its founder Neil J Sloane. The site is a database of over 200000 mathematical sequences, where each sequence is accompanied by a description of the sequence, as well as Mathematica code and in rare instances Maple code used to generate a number of terms of the sequence similar to the graphic below. For this reason, the majority of our research was done in Mathematica.

A004015 Theta series of face-centered cubic (f.c.c.) lattice. 11
(Formerly M4821)
1, 12, 6, 24, 12, 24, 8, 48, 6, 36, 24, 24, 24, 72, 0, 48, 12, 48, 30, 72, 24, 48, 24, 48, 8, 84, 24, 96, 48, 24, 0, 96, 6, 96, 48, 48, 36, 120, 24, 48, 24, 48, 48, 120, 24, 120, 0, 96, 24, 108, 30, 48, 72, 72, 32, 144, 0, 96, 72, 72, 48, 120, 0, 144, 12, 48, 48, 168, 48, 96 (list; graph; refs; listen; history; text; internal format)

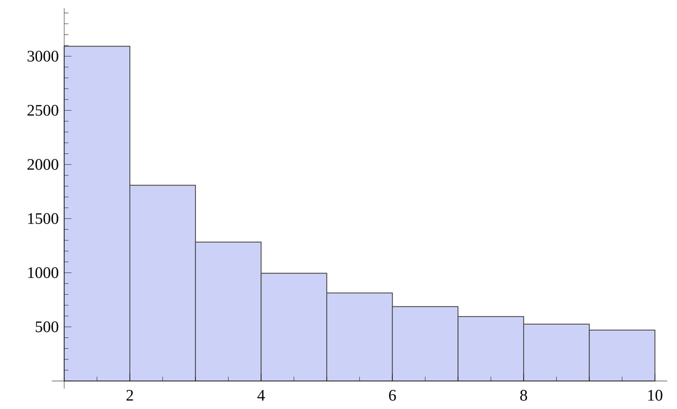
which has a Mathematica formula of $a[n] := SquaresR[3, 2n]; Table[a[n], n, 0, 69]$

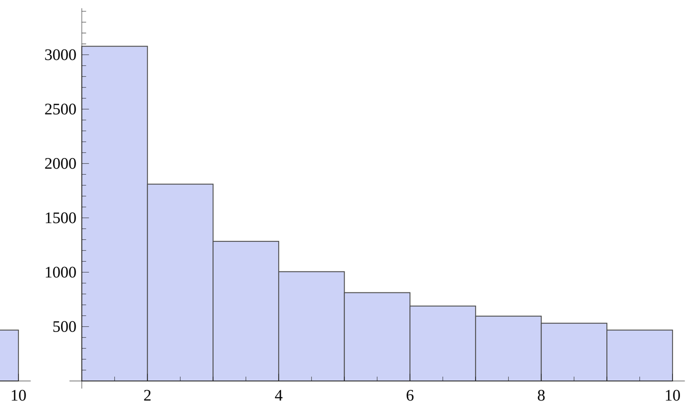
3. Newton's Method

The sequence of errors of iterates generated by Newton's method is modeled by Benford's Law[2]. If f is a polynomial and

$$x_{n+1} = \frac{f(x_n)}{f'(x_n)}$$

is a sequence of Newton's method iterates to a root x_* of f, then the first significant digit of $|x_{n+1} - x_*|$ follows Benford's Law. For each sequence tested, we computed the chi-squared goodness of fit test of the observed significant digits compared to the expected digits from Benford's Law. This sequence had a chi squared value of 0.2295, so we can conclude that the digits likely follow the distribution given by Benford's Law. The histogram for the distribution of leading digits supports this fact. The histogram below on the right is the distribution of the leading digits found in the aforementioned sequence. The histogram on the left is the actual distribution that would be predicted by Benford's Law.



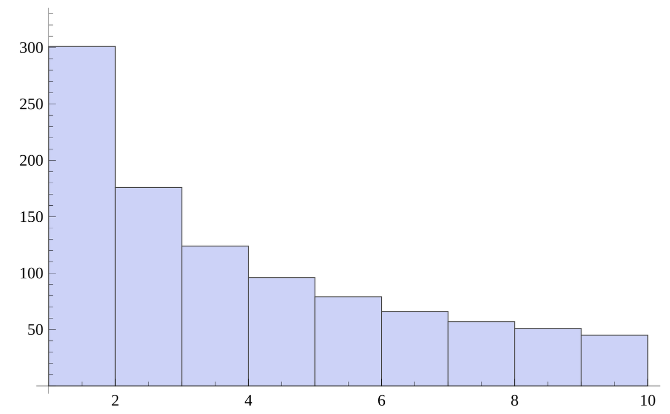


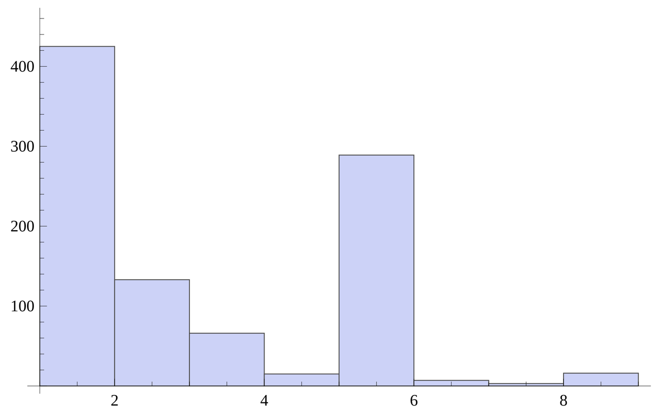
4. Not So Benford Sequences

Over the course of our research, it was apparent that many sequences do not follow Benford's Law, in fact many sequences are very far from being Benford. One such sequence is defined by

$$S_n = \{\text{Number of distinct sets } \{x_i\} | lcm\{x_i\} = n\}$$

For the aforementioned sequence, we obtained a chi-squared value of 844.607. This tells us that the sequence's leading digits have a distribution that is vastly different from that of a Benford sequence. When viewing the histogram of the various digits along with their frequency, it is clear that the sequence is in fact not Benford, as the histogram on the left is for the Benford distribution, while the one on the right depicts our Non-Benford distribution.



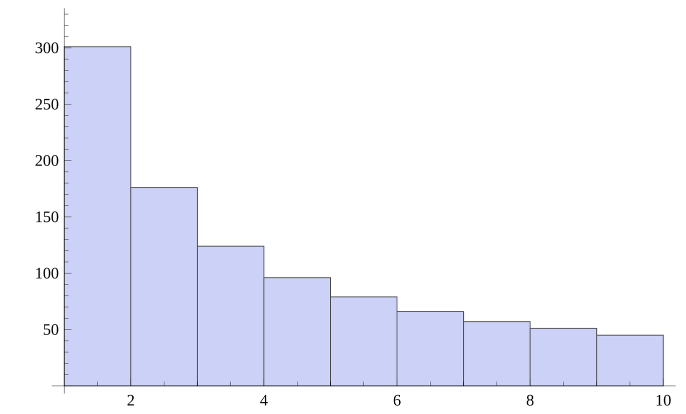


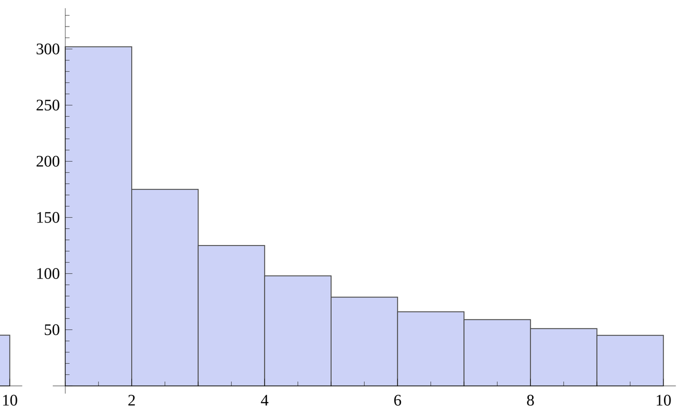
5. Geometric Sequences

Most of the sequences that we analyzed were non-monotonic in order to avoid skewed results. This was due to the fact that the rate of increase for the majority of sequences was slow enough that stopping at a specific term would result in a different distribution of leading digits. However, for certain sequences, the rate of increase was so rapid that ending at one term vs another would alter the distribution very little. For this, we used the geometric sequence $x_n = a * b^n$[5]. One such example of this is the geometric sequence given by:

$$X(n) = 3 \cdot 7^n$$

This produces the following histograms, with the one on the right depicting the leading digits of this geometric sequence, while the one on the right is the histogram for the leading digits under the Benford distribution.





6. Conclusion

Of the many sequences that were tested, there were only a few sequences that follow the distribution of leading digits outlined by Benford's Law. These were:

- Geometric Sequences
- Hailstone Sequences
- Newton's Method Iterates
- Fibonacci Sequences

Currently, however research is being done involving the generalized form of Benford's Law. For $x_n \in [0, N]$

$$P(d) = C \int_d^{d+1} x^{-\alpha(n)} dx = \frac{1}{10^{1-\alpha(n)}}[(d+1)^{1-\alpha(n)} - d^{1-\alpha(n)}]$$

where $\alpha(n) = \frac{1}{log(N)-a}$ where $a = 1.1 \pm .1$. Note that $\alpha = 1$ corresponds to the form of Benford's Law in Box 1. When using the generalized form of Benford's Law, there are a few noteworthy sequences that abide by the distribution given. One of these is the zeros of the Riemann Zeta function. Furthermore, research is being done to show that although a sequence of an infinite number of primes will have leading digits following a uniform distribution, a finite set will follow the Generalized Benford Distribution.[4].

References

[1] The on-line encyclopedia of integer sequences. electronically, 2010.

[2] A. Berger and T. Hill. Newton's method obeys benford's law. *American Mathematical Monthly*, 114(7):588–601, 2007.

[3] Donald Knuth. *The Art of Computer Programming*, volume 2. Addison-Wesley Publishing Company, second edition, 1981.

[4] L. Lacasa and L. Bartolo. The first-digit frequencies of prime numbers and riemann zeta zeros. *Proceedings of The Royal Society for Mathematical, Physical, and Engineering Sciences*, 2009.

[5] S.J. Miller and R. Takloo-bighash. *An invitation to modern number theory*. Number v. 13 in An Invitation to Modern Number Theory. PRINCETON University Press, 2006.

Acknowledgments

- The University of Wisconsin-Eau Claire Mathematics Department.
- The University of Wisconsin-Eau Claire Office of Research and Sponsored Programs.
- Graphics computed and rendered with Mathematica®.
- The University of Wisconsin-Eau Claire Summer Research Experiences for Undergraduates Program