When Life Gives You Lemons - A Statistical Model for Benford's Law

Joseph R. Iafrate '14, Steven J. Miller, Frederick W. Strauch

jri1@williams.edu, sjm1@williams.edu, fws1@williams.edu

Abstract

We attempt to recover the results of Lemons for a probabilistic model of partitioning a conserved quantity. We derive an approximation for the number of partitions of integer X into parts given by a set H as well as the average number of parts $\langle n_j \rangle$ of a given part size x_j.

1. What is Benford's Law?

Benford's Law of Digit Bias concerns the distribution of the leading digits of the elements of a data set. A data set is *Benford* if the probability that the first digit of a set element is d is given by $\log_{10}\left(1+\frac{1}{d}\right)$.

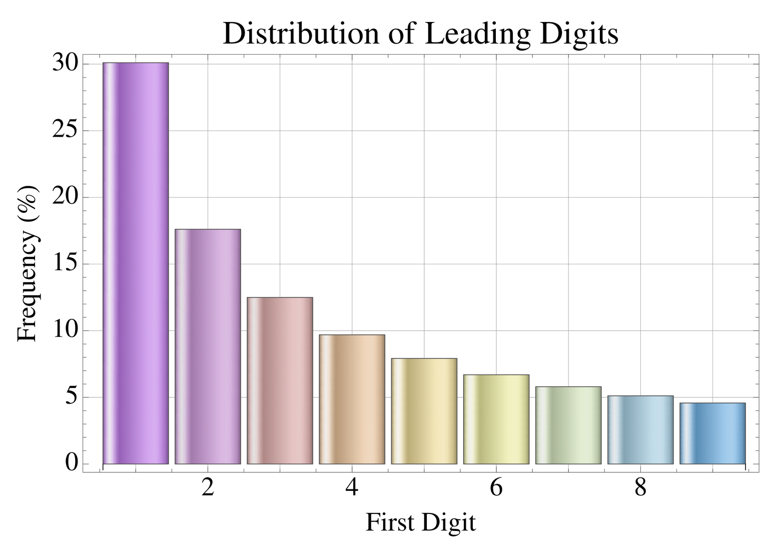

Figure 1: *The distribution of first digits in a data set exhibiting Benford behavior.*

Benford's Law characterizes a large number of real-life data sets. These range from mathematical sets, such as the Fibonacci numbers, to seemingly random collections of numbers, such as the populations of U.S. cities. Knowledge of this distribution can even be used to detect tax or bank fraud.

2. The Problem

Lemons presented the problem in the following way:

- Consider some conserved quantity X.
- Fragment X into $n_j \Delta x_j$ pieces, with sizes between x_j and $x_j + \Delta x_j$, $j = \{1, 2, ..., N\}$.
- Specify the smallest and largest piece sizes and make N divisions between them, such that $X = \sum_{j=1}^{N} x_j n_j \Delta x_j$ (n_j not necessarily an integer).

Using a few theorems regarding averages of constants and random variables, Lemons determined that the average number of pieces of size x_j is given by

$$\langle n_j \rangle = \frac{X}{N x_j \Delta x_j}. \tag{1}$$

Such a distribution, with piece frequency inversely proportional to piece size, leads directly to Benford's Law. Lemons claims that, on average, any fragmentation of a conserved quantity will be Benford.

Our concerns:

- Lemons' definition of piece size is unclear and not practical for models of fragmentation in which pieces are multiples of some base unit (e.g. nucleons in a nuclear fragmentation).
- This result may only hold in particular parameter regimes or under certain assumptions.

Our task is to reproduce Lemons' result for integer pieces.

3. Partitions and $\langle n_j \rangle$

We will cast our model as a restricted partitioning problem:

◇ *Consider an integer X and a set of integers $H = \{x_1, x_2, ..., x_N\}$, $x_j > x_k$ for $j > k$. Let $x_1 = 1$ so that a partition exists for every X.*
◇ *Partition X into integer parts given by H.*
◇ ***What is the average number of parts of size x_j?***

$P(``H", X)$ denotes the number of ways to partition the integer $X$ into parts from $H$. The generating function $\sum_{k=0}^{\infty} P(``H", k) q^k = \prod_{h \in H} (1-q^h)^{-1}$ yields

$$P(``H", X) = \frac{1}{X!} \times \left(\frac{\partial}{\partial q}\right)^X \prod_{h \in H} (1-q^h)^{-1} \big|_{q=0}. \tag{2}$$

When X becomes large, we cannot calculate this. However, as the average number of pieces of size x_j is simply $\frac{total\ n_j\ in\ all\ partitions}{total\ possible\ partitions}$, we *can* use the above to obtain

$$\langle n_j \rangle = \frac{1}{P(``H", X)} \sum_{i=1}^{\lfloor X/x_j \rfloor} P(``H", X - i\, x_j). \tag{3}$$

While in principle this gives us the quantity we desire, we do not have a closed expression for $P(``H", X)$, so we cannot evaluate.

4. First Approach: Approximating $P(``H", X)$

Luckily, we have another definition for $P(``H", X)$:

$$P(``H", X) = \sum_{n_N=0}^{\lfloor \tilde{L}_N \rfloor} \sum_{n_{N-1}=0}^{\lfloor \tilde{L}_{N-1} \rfloor} \dots \sum_{n_2=0}^{\lfloor \tilde{L}_2 \rfloor} \delta\left(X, \sum_{h \in H} n_h x_h\right), \tag{4}$$

where $\tilde{L}_j = \frac{X - \sum_{i=j+1}^{N} n_i x_i}{x_j}$ for $j = \{2, 3, ..., N-1\}$ and $\tilde{L}_N = \frac{X}{x_N}$. We can use the above to obtain an approximate expression for $P(``H", X)$, making the following assumptions/approximations:

- Remove the floors from the $\tilde{L}_j$ but work as if they are integers.
- Assume $X >> N$ and keep only the two highest-order terms of each sum.

Once the dust has cleared, we have

$$P(``H", X) \approx \frac{X^{N-1}}{(N-1)! D_N} + \frac{X^{N-2}}{2(N-2)! D_N}\left(x_2 + \sum_{j=2}^{N} x_j\right), \tag{5}$$

with $D_N = \prod_{i=1}^{N} x_i$. We can plug Equation 5 into Equation 3. In the case that $x_j \mid X$,

$$\langle n_j \rangle = \frac{X}{N x_j}\left(1 + \mathcal{O}\left(\frac{N \langle x \rangle}{X}\right)\right). \tag{6}$$

For a fixed N, as X gets large, the error term becomes insignificant and we recover Lemons' result, albeit only for integer pieces.

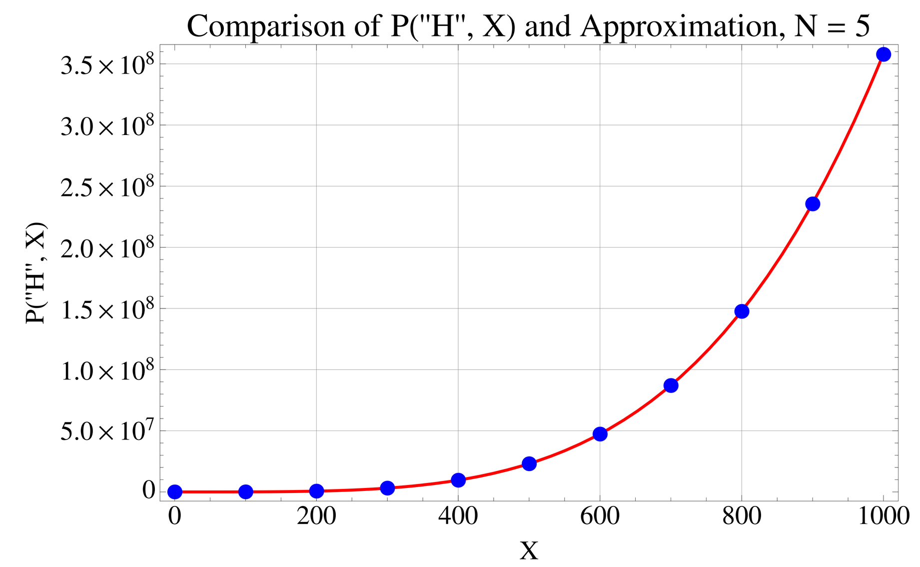

Figure 2: *The points represent the number of partitions of the integer X into parts from the set $H = \{1, 2, 3, 4, 5\}$. The solid line is given by Equation 5.*

5. Second Approach: Bounding $P(``H", X)$

Rather than directly calculate $P(``H", X)$, we might find an upper and lower bound instead. We can use these to bound $\langle n_j \rangle$. We have attempted two methods:

1. Evaluate the sums directly. At each step, replace the L_j with either $\tilde{L}_j$ (to maximize a term) or $\tilde{L}_j - 1$ (to minimize a term) before moving on to the next sum. While calculating an upper bound is fairly straightforward, we have not been able to obtain a lower bound in this way.
2. Replace the sums with integrals and modify the bounds of integration slightly to achieve either an upper or lower bound.

We are currently working to refine the latter method.

6. Future Work

- Finish obtaining bounds for $P(``H", X)$.
- Calculate the variance of n_j. We have already derived an expression for $\langle n_j^2 \rangle$.
- Investigate other fragmentation models, both statistical and stochastic. For these other models, we will pay special attention to the distribution of the first digit of the fragment sizes and identify those circumstances which lead to Benford behavior.

7. Acknowledgements

The authors would like to thank Williams College, whose support made possible this endeavor. This work was funded by NSF Grant DMS0970067.

References

[1] Lemons, Don S. "On the numbers of things and the distribution of first digits." *American Journal of Physics.* 54.9 (1986): 816-17. Web. 29 Jul. 2013.

[2] Andrews, George E. The Theory of Partitions. Reading, Massachusetts: Addison-Wesley Publishing Company, 1976. Print.

[3] Haiman, Mark. "Notes on Partitions and their Generating Functions." . University of California, Berkeley, n.d. Web. 29 Jul 2013. <http://math.berkeley.edu/ mhaiman/math172-spring10/partitions.pdf>.

Summer Research Poster Session, Williams College, August 2, 2013