Benford's Law

VE401 Probabilistic Methods in Engineering
Project 1 (Summer 2017)

Siddharth Ramesh, Ziwen Lu, Kevin Zheng, Derek Tan, Chengcheng Zhu

June 26, 2017

Contents

I Abstract **1**

II Scale Invariance **1**

III Data Visualization **1**

IV Pinkham's Proof **4**

V Benford's Law with Non-leading Digits **6**

VI Shortcomings of Pinkham's Approach **6**

A Appendix **8**

A.I Elastic Properties of the Elements [4] 8

I Abstract

Benford's law [1] is an interesting generalization on how the frequency of the leading digits behave in real life. Intuitively, we would think that, with a large dataset, the distribution of the leading digits would be uniform, 1/9 for every digit $\in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. However, as seen in practice, this is not always the case. This report will explore Benford's law and its property of scale invariance, a proof written by Roger Pinkham [3], as well as the shortcomings of Pinkham's proof as expressed in a proof by Theodore Hill [2].

II Scale Invariance

Let's first investigate the property of re-scaling; this implies that the distribution of the leading digit cannot be uniform.

Let X be a uniformly distributed discrete random variable that corresponds to the first digit, where each digit occurs with probability $\frac{1}{9}$. Assume that, after scaling, then this distribution is independent of re-scaling, meaning that the probability of each digit being a leading digit is still $\frac{1}{9}$.

We denote $P[X = m]$ to mean the probability that the first digit is m, where $m \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, and $P_\alpha[X = m]$ to mean the probability that the first digit is m, after rescaling to a nonzero factor α. By our assumption,

$$P[X = m] = P_\alpha[X = m].$$

By observation, if a number starts with 5, 6, 7, 8, 9, then scaling with a factor $\alpha = 2$ will cause the number to begin with 1. This means that

$$P_2[X = 1] \geq P[X = 5 \cap 6 \cap 7 \cap 8 \cap 9] = \frac{5}{9} \neq P[X = 1]$$

Thus, there is a contradiction between (1) and (2), meaning that the discrete uniformly distributed variable X is not independent of rescaling in all cases.

Benford's law is remarkable in that it applies in a wide variety of natural numbers, such as populations, lengths of rivers, mathematical constants, etc. We have analyzed the periodic tables of elements to show this.

III Data Visualization

We will examine Benford's law in practice through a natural data set, such as the Elastic Properties of the Elements [4]. Below, we have plotted the frequencies of the shear modulus of every element, as well as the frequencies of their leading numbers. A table of the raw data can be found in Appendix A. Visually, the data in Figure 1 seems to be distributed expontentially, with the elements' shear modulus skewed towards lower values.

Figure 1: Frequencies of the Shear Modulus of the Elements in GPa [4]

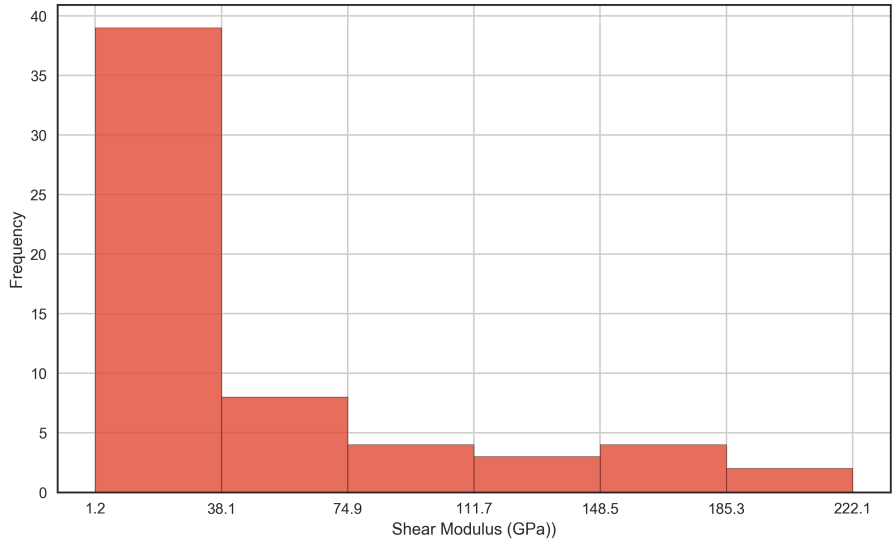
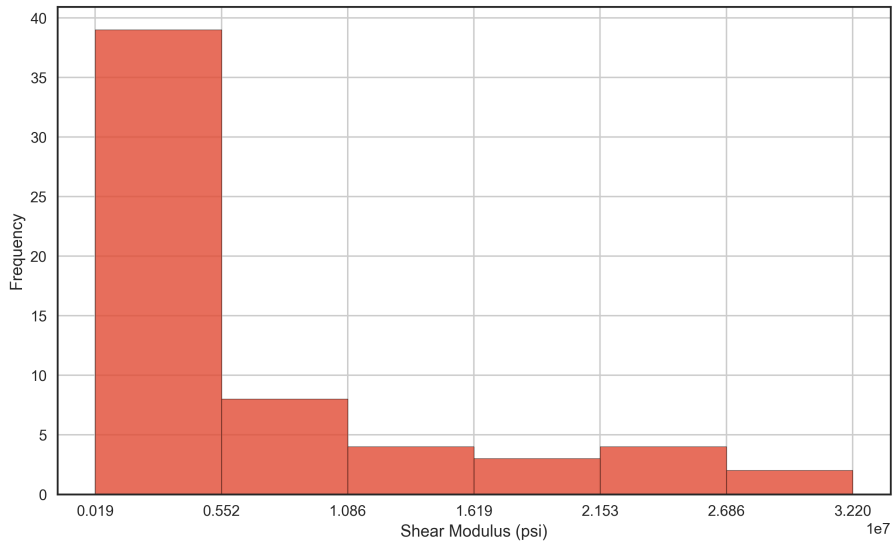


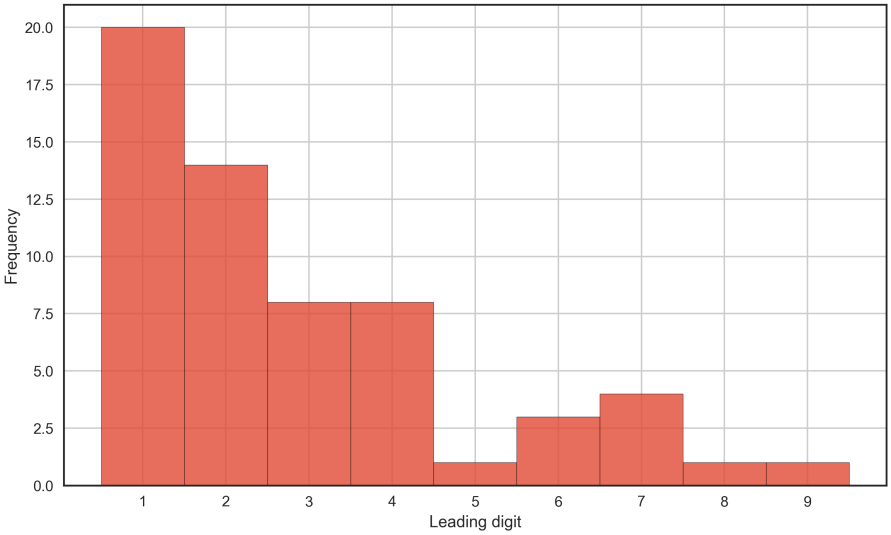
Figure 2: Frequencies of the Shear Modulus of the Elements in psi [4]



The Shear Modulus in psi was calculated by multiplying the Shear Modulus in GPa by a factor of 145038, as 1 gigapascal is 145038 pounds per square inch. Interestingly, the histograms look identical, with identical frequencies, since the category length between the bins have been scaled by the same amount. This shows that changing a datasets units by a multiplicative factor would not affect the histogram of the frequencies of the data, as long as the bin length is scaled by the same factor.
Now, let us see if the leading digits follow Benford's law.

According to Figure 3, it can be seen that 1 is the most common leading digit, 2 is

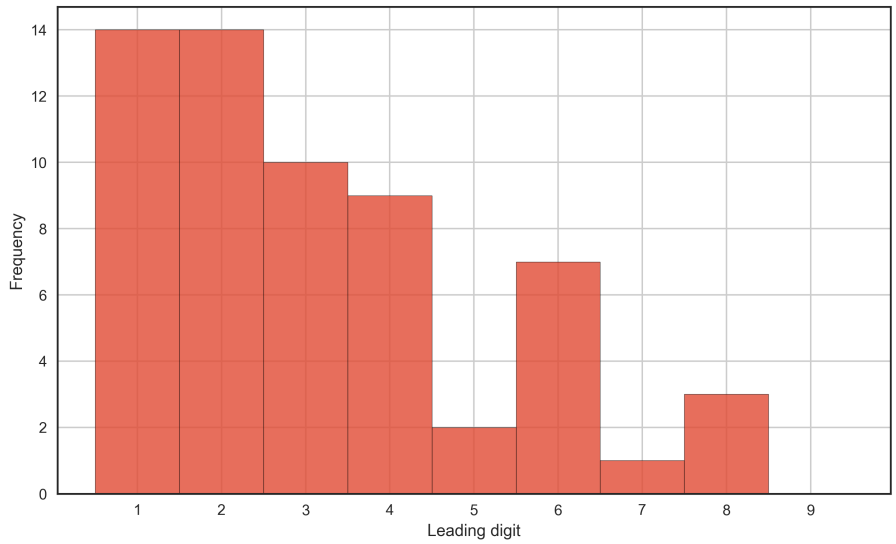
Figure 3: Leading digits of the Shear Modulus of the Elements in GPa [4]



second common, and the frequencies of the remaining digits decrease. Note that these do not follow Benford's distribution perfectly; there are more digits who lead with 6, 7 rather than 5, 8, 9. However, the frequencies are clearly not uniform.

To see the scale invariance of Benford's law, the leading digits of the scaled values to psi are plotted in Figure 4. From Figure 4, we can see that there is a similar trend that the leading digits tend to cluster towards the left. Note that this trend is not perfect; it does not mirror the leading digits of the GPa dataset in the same way as the histograms do for the raw frequencies.

Figure 4: Leading digits of the Shear Modulus of the Elements in psi [4]



IV Pinkham's Proof

Pinkham [3] provides a systematic proof of Benfords law. For a random set of data without physical constraint, for example the Youngs modulus of a given set of metals, F(x) denotes the ratio of data less than x. For example, F(10) refers to the ratio of the amount of data less than 10 to the total amount of data. Then, the ratio of data with the first leading digit starting from 1 to x-1 is given by:

$$D(x) = \sum_{m=-\infty}^{\infty} [F(x10^m) - F(10^m)] \tag{1}$$

Let us look at this using an example. Suppose x=2, then the summation on the right sums over the ratio of data between 0.1-0.2, 10-20, 100-200, 1000-2000, ..., all of which start with the digit 1. In other words, the summation gives the ratio of data with the first leading digit to be 1. For the same reason, D(3) gives the probability of finding a number starting with 1 or 2, so it is cumulative.

Later on, Pinkham introduced the invariance principle, which essentially says that the distribution of leading digits does not depend on the units of measurement. A very straightforward example would be a list of river lengths in a specific country. The ratio of the amount of data with the leading digit 1 is the same whether the units of measurement is in metres or kilometres. If we were to introduce a constant α as the conversion factor, then the invariance principle can be defined by:

$$\begin{aligned} D(x) &= \sum_{m=-\infty}^{\infty} [F(x10^m) - F(10^m)] && x \in \{1,2,3,4,5,6,7,8,9\} \\ &= \sum_{m=-\infty}^{\infty} [F(\frac{x}{\alpha}10^m) - F(\frac{1}{\alpha}10^m)] && (2) \end{aligned}$$

D(x) is still the ratio of numbers with the leading digits from 1 to x-1. To avoid making it too complicated, let us stick with the river length example. In this case, the conversion factor is $\alpha = 1000$. X can be any random integer from 2 to 9. Lets assume X=3. What this equation means is that the ratio of data with leading digits 1 or 2 in the unit of meters is the same as the ratio of data with leading digit 1 or 2 in the unit of kilometers, but the corresponding value in kilometers is 1000 times smaller. Or one can understand the equation in this way: the ratio of river with length 1000-3000 meters long is the same as the ratio of river with length 1km-3km long.

$$D(\frac{x}{\alpha}) = \sum_{m=-\infty}^{\infty} [F(\frac{x}{\alpha}10^m) - F(10^m)] \qquad (\alpha > 0) \quad (3)$$

$$D(\frac{1}{\alpha}) = \sum_{m=-\infty}^{\infty} [F(\frac{1}{\alpha}10^m) - F(10^m)] \qquad (\alpha > 0) \quad (4)$$

$$\begin{aligned}(3) - (4) &= \sum_{m=-\infty}^{\infty} [F(\frac{x}{\alpha}10^m) - F(10^m)] - \sum_{m=-\infty}^{\infty} [F(\frac{1}{\alpha}10^m) - F(10^m)] \\ &= \sum_{m=-\infty}^{\infty} F(\frac{x}{\alpha}10^m) - F(\frac{1}{\alpha}10^m) = D(x) \\ &\therefore \; D(\frac{x}{\alpha}) - D(\frac{1}{\alpha}) = D(x)\end{aligned}$$

Equation 5 shows that the function D has a similar property with the function $log_{10}(x)$. Therefore, we can deduce that $D(x) = log_{10}(x)$, which tells us that the ratio of data with the leading digit from 1 to x-1 is equal to $log_{10}(x)$. Furthermore, the ratio of data starting with digit x equals to $log(x + 1) - log(x)$.

V Benford's Law with Non-leading Digits

As observed in the previous sections, Benford's distribution models the probability of a digit occurring as the first digit of a sequence in naturally occurring systems. The probability of this distribution [1] is modelled by

$$P = log(n+1) - log(n)$$

So, the probability of the number 1 being the first digit of a sequence is log(2) - log(1). However, this formula can be extended to other non-leading digits in the sequence.
If the probability of a number occurring as the first digit is P = log(n+1)-log(n), then the probability of the first two numbers of the sequence being 10 is log(11) - log(10). Therefore, if we wanted to find the probability that the second digit of a sequence was 0, we would have to add up the probabilities of the cases in which this is possible (10, 20, 30, 40, 50, 60, 70, 80 and 90). So

$$\begin{aligned} P(\text{second digit} = 0) &= log(11) - log(10) + log(21) - log(20) + ... + log(91) - log(90) \\ &\approx 0.119 = 11.9\% \end{aligned}$$

Similarly, the probability that the second digit will be a 1 is 0.114, the second digit being a 2 is 0.109, and so on. This is how we can use Benford's law to predict the second digit. We can already see that the probability that the second digit will be a 0 is greater than the probability the second digit will be a 1, but it can also be seen that while the probabilities of greater numbers occurring is still less, the difference is now significantly lower. This difference continually decreases as we increases the order of decimal places, until we reach the 9th decimal place, when the probability of all number occurring is equal, at 0.1.

VI Shortcomings of Pinkham's Approach

Theodore Hill [2] provided several counter arguments to the common approaches used by several authors to prove Benford's Law. Specifically, these approaches included discrete, continous methods, as well as the hypothesis of scale-invariance which Pinkham's approach focused on.
For the discrete methods, proofs have made the assumption that the data set is based on the natural numbers, so some density function mapping to a set of numbers $F_d = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ would be natural. Hill shows this to be false - that this set F_d is not in fact a natural density. Further, Benford's data included continous data (such as irrational numbers), which is a significant oversight for a discrete argument. Additionally, for both the discrete and continuous methods, Hill points out that the proofs only prove the finite additivity axiom; that is,

$$P\big[\bigcup_{n=1}^{N} = A_n\big] = \sum_{n=1}^{N} P\big[A_n\big]$$

They do not, however, satisfy the countable additivity axiom:

$$P\big[\bigcup_{n=1}^{\infty} = A_n\big] = \sum_{n=1}^{\infty} P\big[A_n\big]$$

thus, it is a weaker proof.
The final method is the proof under the assumption that Benford's law would be scale invariant. For something to be scale invariant, it means that for a random variable X, that its distribution is the same as αX. The only way for this random variable to be scale invariant is for it to be zero, so that $P[X = 0] = 1$. This is the only random variable that has this property, as [5]

$$\begin{aligned} P[X \neq 0] &= lim_{a\to+\infty} P[|X| > a^{-1}] \\ &= lim_{a\to+\infty} P[|a^2 X| > a] \\ &= lim_{a\to+\infty} P[|X| > a] = 0 \end{aligned} \tag{5}$$

Thus, since the set F_d does not include zero, the random variable on this density can't be zero, and can't be scale invariant. Additionally, Hill adds that the scale invariance hypothesis has the same problems as the former two methods, so it cannot be strong enough to be a thorough proof.

A Appendix

A.I Elastic Properties of the Elements [4]

Atomic Number	Symbol	Name	Shear Modulus (GPa)	Shear Modulus (psi)	Leading Digit (GPa)	Leading Digit (psi)
3	Li	lithium	4.2	609159.6	4	6
4	Be	beryllium	132.0	19145016.0	1	1
11	Na	sodium	3.3	478625.4	3	4
12	Mg	magnesium	17.0	2465646.0	1	2
13	Al	aluminium	26.0	3770988.0	2	3
19	K	potassium	1.3	188549.4	1	1
20	Ca	calcium	7.4	1073281.2	7	1
21	Sc	scandium	29.1	4220605.8	2	4
22	Ti	titanium	44.0	6381672.0	4	6
23	V	vanadium	47.0	6816786.0	4	6
24	Cr	chromium	115.0	16679370.0	1	1
26	Fe	iron	82.0	11893116.0	8	1
27	Co	cobalt	75.0	10877850.0	7	1
28	Ni	nickel	76.0	11022888.0	7	1
29	Cu	copper	48.0	6961824.0	4	6
30	Zn	zinc	43.0	6236634.0	4	6
34	Se	selenium	3.7	536640.6	3	5
38	Sr	strontium	6.1	884731.8	6	8
39	Y	yttrium	25.6	3712972.8	2	3
40	Zr	zirconium	33.0	4786254.0	3	4
41	Nb	niobium	38.0	5511444.0	3	5
42	Mo	molybdenum	120.0	17404560.0	1	1
44	Ru	ruthenium	173.0	25091574.0	1	2
45	Rh	rhodium	150.0	21755700.0	1	2
46	Pd	palladium	44.0	6381672.0	4	6
47	Ag	silver	30.0	4351140.0	3	4
48	Cd	cadmium	19.0	2755722.0	1	2
50	Sn	tin	18.0	2610684.0	1	2
51	Sb	antimony	20.0	2900760.0	2	2
52	Te	tellurium	16.0	2320608.0	1	2
56	Ba	barium	4.9	710686.2	4	7
57	La	lanthanum	14.3	2074043.4	1	2
58	Ce	cerium	13.5	1958013.0	1	1
59	Pr	praseodymium	14.8	2146562.4	1	2
60	Nd	neodymium	16.3	2364119.4	1	2
61	Pm	promethium	18.0	2610684.0	1	2
62	Sm	samarium	19.5	2828241.0	1	2

Atomic Number	Symbol	Name	Shear Modulus (GPa)	Shear Modulus (psi)	Leading Digit (GPa)	Leading Digit (psi)
63	Eu	europium	7.9	1145800.2	7	1
64	Gd	gadolinium	21.8	3161828.4	2	3
65	Tb	terbium	22.1	3205339.8	2	3
66	Dy	dysprosium	24.7	3582438.6	2	3
67	Ho	holmium	26.3	3814499.4	2	3
68	Er	erbium	28.3	4104575.4	2	4
69	Tm	thulium	30.5	4423659.0	3	4
70	Yb	ytterbium	9.9	1435876.2	9	1
71	Lu	lutetium	27.2	3945033.6	2	3
72	Hf	hafnium	30.0	4351140.0	3	4
73	Ta	tantalum	69.0	10007622.0	6	1
74	W	tungsten	161.0	23351118.0	1	2
75	Re	rhenium	178.0	25816764.0	1	2
76	Os	osmium	222.0	32198436.0	2	3
77	Ir	iridium	210.0	30457980.0	2	3
78	Pt	platinum	61.0	8847318.0	6	8
79	Au	gold	27.0	3916026.0	2	3
81	Tl	thallium	2.8	406106.4	2	4
82	Pb	lead	5.6	812212.8	5	8
83	Bi	bismuth	12.0	1740456.0	1	1
90	Th	thorium	31.0	4496178.0	3	4
92	U	uranium	111.0	16099218.0	1	1
94	Pu	plutonium	43.0	6236634.0	4	6

References

[1] Frank Benford. The law of anomalous numbers. Proc. Amer. Philosophical Soc., 78:551572, 1938. http://www.jstor.org/stable/984802.

[2] Theodore P. Hill Base-invariance implies Benfords law. Proc. Amer. Math. Soc., 123:887895, 1995. http://www.ams.org/journals/proc/1995-123-03/S0002-9939-1995-1233974-8/.

[3] Roger S. Pinkham. On the distribution of first significant digits. Ann. Math. Statist., 32(4):12231230, 12

[4] Wikipedia. Elastic properties of the elements (data page) wikipedia, the free encyclopedia, 2013. https://en.wikipedia.org/w/index.php?title=Elastic_properties_of_the_elements_(data_page) Web. Accessed June 29th, 2015.

[5] Berger, Arno, and Theodore P. Hill. An Introduction to Benford's Law. PRINCETON; OXFORD, Princeton University Press, 2015. JSTOR, www.jstor.org/stable/j.ctt1dr35m0.