UNIVERSITY OF MICHIGAN

JOINT INSTITUTE
交大密西根学院

PROBABILISTIC METHODS IN ENGINEERING
(VE-401)

TERM PROJECT REPORT 1

2017 Spring Team 14

Name:	Chen Youjie	ID:	5143709126
	Fan Yuanzhen		515370910067
	Kang Yiyan		5143709052
	Shen Chunhao		515370910210
	Ye Jiaxin		515370910209

Instructor:	Horst Hohberger
Teaching Assistant:	Wang Zesen
	Zhu Weiyu

Date: Apirl 11, 2017

Contents

Abstract

In this report, we will mainly focus on Benford's Distribution and its property. First, we will offer a proof for the dependency between re-scaling and discrete uniform distribution. Second, to further prove this statement, we will use Elastic Properties of Elements [1] to get the result. At the same time, we transform the unit of the original data from GPa to psi for recalculation. Third, we restate Pinkham's proof [2] that scale invariance implies Bendord's law. In addition, we investigate the frequency of higher-order decimal digits. Finally, we have some further considerations about Benford's law, which is the completeness of its proof.

1 Objectives

This project is aimed at applying our basic knowledge of probabilistic methods in engineering including probability and statistics to several interesting issues beyond class assignments. It requires our team efforts on each question, involing comprehension, data collection, discussion and, more importantly, the balence between individual thoughts and team collaboration.

2 Re-scaling of Discrete Uniform Distribution

In this part, we will try to prove the following statement:

If the leading digits of a discrete random variable follow a discrete uniform distribution, then this distribution is not independent of re-scaling.

We prove this by contradiction.

First, we assume the statement, at the moment, to be true. Then, we define a sample space given by

$$A = \{x|1 \leq x < 10, x \in N^*\} \tag{2.1}$$

The leading digit of each sample point in this sample space has probability of $\frac{1}{9}$ to occour; therefore, the leading digit forms a random variable, $(X,\, f_X)$, where$X \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ and $f_X = \frac{1}{9}$.

Since the probabilities of each leading number are the same, they are uniformly distributed.

Now, we define another two sample spaces called $A'$and A'', which stands for the two new random variables, $(Y,\, f_Y)$ and $(Z,\, f_Z)$ from diving X by 2 and 3, respectively.

In detail,

$$A' = \{y|\frac{1}{2} \leq y \leq \frac{9}{2}, y \in Q\} \tag{2.2}$$

$$f'_1 = f'_2 = f'_3 = f'_4 = \frac{2}{9} \tag{2.3}$$

$$f'_5 = \frac{1}{9} \tag{2.4}$$

$$f'_6 = f'_7 = f'_8 = f'_9 = 0 \tag{2.5}$$

$$A'' = \{z|\frac{1}{3} \leq z \leq 3, z \in Q\} \tag{2.6}$$

$$f''_1 = f''_2 = \frac{1}{3} \tag{2.7}$$

$$f''_3 = \frac{2}{9} \tag{2.8}$$

$$f''_6 = \frac{1}{9} \tag{2.9}$$

$$f''_4 = f''_5 = f''_7 = f''_8 = f''_9 = 0 \tag{2.10}$$

Obviously, the density function of leading digits are not uniform after the re-scaling. At the same time, the density function of the leading numbers are different between re-scaling by 2 and 3. This contradicts with our previous assumption.

Therefore, we can easily draw the conclusion that if the leading digits of a discrete random variable follow a discrete uniform distribution, then this distribution is not independent of re-scaling.

3 Frequency of Leading Digits of Shear Modulus

Consider that we have learned much about the scaling argument, we would like to use an example to illustrate how the leading digits of the data distribute and how their distribution becomes after re-scalling.

3.1 Shear Modulus data with GPa Unit

We choose to analyze the values of shear modulus of several solid elements. Shear modulus is a mechanical property that measure the shear deformation of a material corresponding to shear stress. We use the data of Elastic properties of the elements [1] and have included it in the report as Table 1.

Table 1: Shear Modulus of Several elements (GPa)

number	symbol	name	use	WEL	CRC
3	Li	lithium	4.2	4.2	
4	Be	beryllium	132	132	
11	Na	sodium	3.3	3.3	
12	Mg	magnesium	17	17	
13	Al	aluminium	26	26	
19	K	potassium	1.3	1.3	
20	Ca	calcium	7.4	7.4	
21	Sc	scandium	29.1	29	29.1

number	symbol	name	use	WEL	CRC
22	Ti	titanium	44	44	
23	V	vanadium	47	47	
24	Cr	chromium	115	115	
26	Fe	iron	82	82	
27	Co	cobalt	75	75	
28	Ni	nickel	76	76	
29	Cu	copper	48	48	
30	Zn	zinc	43	43	
34	Se	selenium	3.7	3.7	
38	Sr	strontium	6.1	6.1	
39	Y	yttrium	25.6	26	25.6
40	Zr	zirconium	33	33	
41	Nb	niobium	38	38	
42	Mo	molybdenum	120	120	
44	Ru	ruthenium	173	173	
45	Rh	rhodium	150	150	
46	Pd	palladium	44	44	
47	Ag	silver	30	30	
48	Cd	cadmium	19	19	
50	Sn	tin	18	18	
51	Sb	antimony	20	20	
52	Te	tellurium	16	16	
56	Ba	barium	4.9	4.9	
57	La	lanthanum	(α form) 14.3	14	(α form) 14.3
58	Ce	cerium	(γ form) 13.5	14	(γ form) 13.5
59	Pr	praseodymium	(α form) 14.8	15	(α form) 14.8
60	Nd	neodymium	(α form) 16.3	16	(α form) 16.3
61	Pm	promethium	(α form) est. 18	18	(α form) est. 18
62	Sm	samarium	(α form) 19.5	20	(α form) 19.5
63	Eu	europium	7.9	7.9	7.9
64	Gd	gadolinium	(α form) 21.8	22	(α form) 21.8
65	Tb	terbium	(α form) 22.1	22	(α form) 22.1
66	Dy	dysprosium	(α form) 24.7	25	(α form) 24.7
67	Ho	holmium	26.3	26	26.3
68	Er	erbium	28.3	28	28.3
69	Tm	thulium	30.5	31	30.5
70	Yb	ytterbium	(β form) 9.9	9.9	(β form) 9.9
71	Lu	lutetium	27.2	27	27.2
72	Hf	hafnium	30	30	

number	symbol	name	use	WEL	CRC
73	Ta	tantalum	69	69	
74	W	tungsten	161	161	
75	Re	rhenium	178	178	
76	Os	osmium	222	222	
77	Ir	iridium	210	210	
78	Pt	platinum	61	61	
79	Au	gold	27	27	
81	Tl	thallium	2.8	2.8	
82	Pb	lead	5.6	5.6	
83	Bi	bismuth	12	12	
90	Th	thorium	31	31	
92	U	uranium	111	111	
94	Pu	plutonium	43	43	

According to the 60 data in Table 1, we record the frequencies that every digit shows up as the leading digit and conclude the result in the Table 2.

Table 2: Frequencies of leading digits for Table 1 in GPa

Leading Digits	1	2	3	4	5	6	7	8	9
Times	16	13	11	9	1	2	6	1	1
Frequencies	0.267	0.217	0.183	0.150	0.017	0.033	0.100	0.017	0.017

As shown in Table 2, 1 appears as the leading digit for 16 times, which is the highest among all nine digits. The relatively large digits, such as 5, 6, 8 and 9, appears much less than 1, 2, 3 and 4. So far, we draw a corresponding histogram of Table 2.

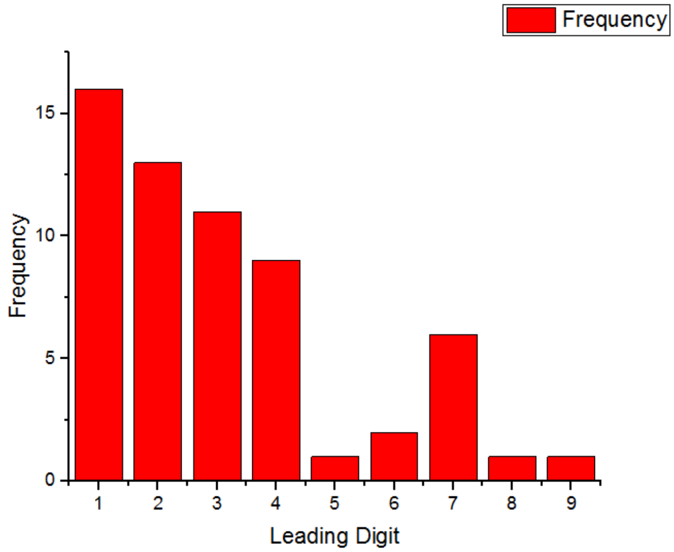

Figure 1: Frequencies of Leading Digits in GPa

In Figure 1, the frequencies decrease as the corresponding leading digits increase, except that the frequency of 7 suddenly increases. This indicates that there is a tendency that small digits are more likely to be the elading digits, for values of shear modulus. However, we attribute the insufficient data to the reason why 7 has higher freqency than 5 or 6. Thus, it infers that by increasing the number of data we collect, the frequency of 7 will follow the tendency we find out previously.

3.2 Shear Modulus data with psi Unit

To check whether the tendency we find in Section 3.1 is reasonable, we recalculate the shear moduli by using psi(pounds per square inch) as the unit. Then we repeat the procedure in Section 3.1 to complete the data result and the histogram.

First, the unit conversion equation is given by

$$1 \text{ psi} = 6894.757 \text{ Pa} \tag{3.1}$$

After converting the units, we can easily transform Table 1 to Table 3.

Table 3: Shear Modulus of Several elements (psi)

number	symbol	name	use
3	Li	lithium	609158.52436858905
4	Be	beryllium	19144982.194441371
11	Na	sodium	478624.55486103427
12	Mg	magnesium	2465641.6462538131
13	Al	aluminium	3770981.3413293608
19	K	potassium	188549.06706646807
20	Ca	calcium	1073279.3048398951
21	Sc	scandium	4220598.3474109387
22	Ti	titanium	6381660.7314804569
23	V	vanadium	6816773.9631723063
24	Cr	chromium	16679340.548187558
26	Fe	iron	11893094.999577215
27	Co	cobalt	10877830.792296235
28	Ni	nickel	11022868.536193516
29	Cu	copper	6961811.7070695898
30	Zn	zinc	6236622.9875831744
34	Se	selenium	536639.65241994755
38	Sr	strontium	884730.23777342704
39	Y	yttrium	3712966.243770448
40	Zr	zirconium	4786245.5486103427
41	Nb	niobium	5511434.2680967581
42	Mo	molybdenum	17404529.267673973
44	Ru	ruthenium	25091529.694229979
45	Rh	rhodium	21755661.584592469
46	Pd	palladium	6381660.7314804569
47	Ag	silver	4351132.3169184932
48	Cd	cadmium	2755717.134048379
50	Sn	tin	2610679.390151096
51	Sb	antimony	2900754.8779456625
52	Te	tellurium	2320603.9023565301
56	Ba	barium	710684.9450966873
57	La	lanthanum	2074039.7377311485
58	Ce	cerium	1958009.5426133221
59	Pr	praseodymium	2146558.6096797902
60	Nd	neodymium	2364115.2255257149
61	Pm	promethium	2610679.390151096
62	Sm	samarium	2828236.0059970208
63	Eu	europium	1145798.1767885366

number	symbol	name	use
64	Gd	gadolinium	3161822.816960772
65	Tb	terbium	3205334.1401299569
66	Dy	dysprosium	3582432.274262893
67	Ho	holmium	3814492.6644985462
68	Er	erbium	4104568.1522931121
69	Tm	thulium	4423651.188867135
70	Yb	ytterbium	1435873.6645831028
71	Lu	lutetium	3945026.6340061007
72	Hf	hafnium	4351132.3169184932
73	Ta	tantalum	10007604.328912536
74	W	tungsten	23351076.767462581
75	Re	rhenium	25816718.413716394
76	Os	osmium	32198379.145196851
77	Ir	iridium	30457926.218429454
78	Pt	platinum	8847302.3777342699
79	Au	gold	3916019.0852266443
81	Tl	thallium	406105.68291239272
82	Pb	lead	812211.36582478543
83	Bi	bismuth	1740452.9267673974
90	Th	thorium	4496170.0608157767
92	U	uranium	16099189.572598426
94	Pu	plutonium	6236622.9875831744

Similarly, we record the frequencies that every digit shows up as the leading digit and conclude the result in the Table 4.

Table 4: Frequencies of leading digits for Table 3 in psi

Leading Digits	1	2	3	4	5	6	7	8	9
Times	14	14	10	9	2	7	1	3	0
Frequencies	0.233	0.233	0.167	0.150	0.033	0.117	0.017	0.050	0.000

As shown in Table 4, both 1 and 2 appear as the leading digit for relatively 14 times, which is the highest among all nine digits. Analogously, 5, 7, 8 and 9, appears much less than 1, 2, 3 and 4. So far, we draw a corresponding histogram of Table 4.

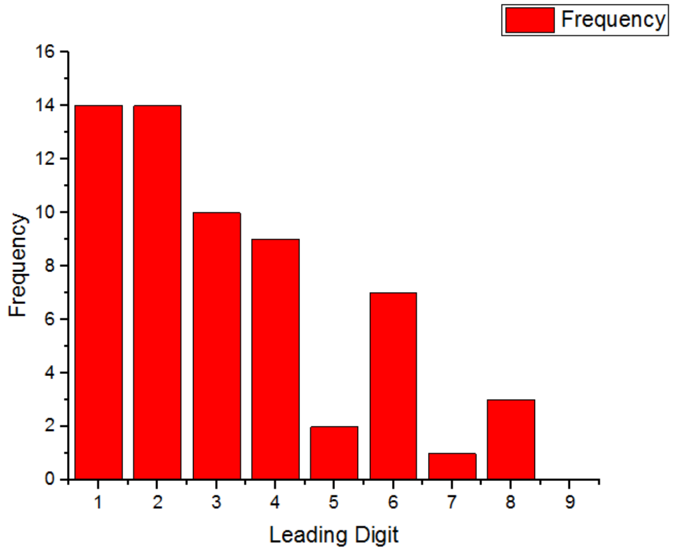

Figure 2: Frequencies of Leading Digits in psi

In Figure 2, the tendency of the frequencies measure in psi are quite similar to those measure in Pa. They decrease as the corresponding leading digits increase, except that this time, the frequency of 6 suddenly increases. This may probably be the result of the increasing 7 in Figure 1.

3.3 Comparison and Conclusion

Comparing the frequencies of shear modulus in both Pa and psi, we can easily reach a conclusion that the distribution of the leading digits are not uniform. The tendency is confirmed that smaller digits' frequencies are higher than larger ones'.

In addition, we have checked the distribution of the leading digits according to the Benford's distribution.

Table 5: Benford's Distribution of Leading Digits

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

By comparing Table 2, Table 4 and Table 5, our data of shear modulus conforms with the probabilities given by Benform distribution. This indicates that the value of shear modulus is natural data.

4 Scaling Argument

4.1 Introduction

Up to now, it is convinced that given a naturally occurring numbers whose sizes are not constrained by outside effects, the distribution of the leading digits should not change after re-scaling. In Section 2, we directly turn down the intuitive thought that the leading digit should follow a uniform distribution by the fact that it does not obey the *scaling argument*. Yet we still have not derived a specific density function for this underlying distribution and in the meanwhile, figured out why it should follow the scaling argument, if $Benford'sLaw$ actually holds. Inspired by the example given in both Section 3.1 and Section 3.2, the cumulative distribution function for the first digit n or less is approximately $log_{10}(n+1)$. In the following discussion, we will first verify this approximation and then prove the validity of the scaling argument.

4.1.1 Pinkham's Way of Verifying the Benford's Law

i) Consider two random variables (Φ, f_Φ) and (Θ, g_Θ), where Φ is determined entirely by Θ with

$$\phi \equiv \theta \ mod\ (1) \qquad 0 \leq \phi < 1, \quad -\infty < \theta < \infty. \tag{4.1}$$

Alternatively we can express ϕ by

$$\phi + m = \theta \qquad -\infty < m < \infty. \tag{4.2}$$

If we take dx and dy to be small, approximately

$$P[x \leq \theta < x + dx] = g(x)dx \tag{4.3}$$

$$P[y \leq \phi < y + dy] = f(y)dy \tag{4.4}$$

The cumulative distribution function $F(y)$ for $f(y)$ is

$$f(y)dy = \sum_{m=-\infty}^{\infty} g(y+m)dy \tag{4.5}$$

$$f(y) = \sum_{m=-\infty}^{\infty} g(y+m) \tag{4.6}$$

For a wide range of possible distributions of θ the distribution of ϕ should be approximately uniform.

$$f(y) = \sum_{m=-\infty}^{\infty} g(y+m) \approx 1, \qquad 0 \leq y \leq 1 \tag{4.7}$$

ii) Based on the above conclusion, we can then verify the correctness of $log_{10}(n+1)$. First, Let F(x) be the cumulative distribution function for the population of physical constants.

$$D(x) = \sum_{m=-\infty}^{\infty} [F(x10^m) - F(10^m)], \qquad x > 0 \tag{4.8}$$

$D(x)$ here is defined to be continuous. $F(x10^m) - F(10^m)$ gives the number between $x10^m$ and 10^m, so if we take $x = 2, ...10$, respectively it gives the proportion of the leading digit $x - 1$ or less.

Then we let $D(x) = log_{10}x$, $y = log_{10}x$, $G(y) = F(10^y)$, we get

$$y \approx \sum_{m=-\infty}^{\infty} [G(y+m) - G(m)] \tag{4.9}$$

take derivatives,

$$1 \approx \sum_{m=-\infty}^{\infty} g(y+m) \tag{4.10}$$

It is consistent with the i), Benford's Law is supported. But later, people find it is not a full proof of Benford's Law, it has short comings, which will be discussed in Section 6.

4.1.2 Theory of Continued Fractions

We give a basic introduction of continued fractions here [3].

A continued fraction is an expression of the form

$$a_0 + \frac{b_1|}{|a_1} + \frac{b_2|}{|a_2} + ... \frac{b_n|}{|a_n} + ... \tag{4.11}$$

where

$$\{a_n\}_{n=0}^{w}, \qquad \{b_n\}_{n=1}^{w}. \tag{4.12}$$

Alternatively, we can also use the notation

$$a_0 + \frac{b_1}{a_1 + \frac{b_2}{a_2 + \frac{b_3}{\cdots}}}. \tag{4.13}$$

Then for every continued fraction the recurrence equation

$$P_n = a_n P_{n-1} + b_n P_{n-2} \tag{4.14}$$

$$Q_n = a_n Q_{n-1} + b_n Q_{n-2} \tag{4.15}$$

$$0 \le n \le \omega + 1 \tag{4.16}$$

with the initial conditions

$$b_0 = 1,\ P_{-2} = 0,\ P_{-1} = 1,\ Q_{-2} = 1,\ Q_{-1} = 0 \tag{4.17}$$

The fraction $\delta = \frac{P_n}{Q_n}$ is called the n-th convergent of the continued fraction.

$$\delta_0 = a_0, \quad \delta_1 = a_0 + \frac{b_1}{a_1}, \quad \delta_2 = a_0 \frac{b_1}{a_1 + \frac{b_2}{a_2}} \quad ... \tag{4.18}$$

4.1.3 Dense Set $\{a_n\}$

This section gives a proof of the following theorem.

Suppose a is irrational, and let [x] denote the largest integer not exceeding x, then the sequence

$$a_n = na - [na], \qquad n = 1, 2, ... \tag{4.19}$$

is uniformly distributed on [0,1].We only to prove that for any $m \in \mathbb{N} \backslash \{0\}$

$$\{a_n\} \bigcap [\frac{k-1}{m}, \frac{k}{m}] \neq \emptyset. \qquad k = 1, 2, 3...m, \qquad m \longrightarrow \infty \tag{4.20}$$

By Dirichlet principle there are two numbers x and y, so a_x and a_y are in the same interval $[\frac{k-1}{m}, \frac{k}{m}]$.

Suppose $x > y$

$$xa = a_x + [xa], \quad ya = a_y + [ya] \tag{4.21}$$

$$(x-y)a = (a_x - a_y) + [xa] - [ya] \tag{4.22}$$

$$a_{x-y} \in [0, \frac{1}{m}] \ \ or \ \ a_{x-y} \in [\frac{m-1}{m}, 1] \tag{4.23}$$

As long as take all multiples of a_{x-y}, then in each of the m intervals must at least one of the values pa_{x-y}, $p \in \mathbb{Z}$.

Thus there exists a subsequence $a_{n'}$ converging to any fixed h($0 \leq h < 1$).

4.2 Scaling Argument

Previously it is verified that the leading digit being n or less follows a distribution approximately $log_{10}(1+n)$. Furthermore, it is convinced that Benford's Law holds because the occurrence digits should follow a distribution that does not change after rescaling.

Prove the sufficiency, which goes if the data satisfies the property that the leading digit distribution does not change after rescaling, the Benford's Law holds, the distribution is $log_{10}(1+n)$ for n or less.

Before rescaling, let $F(x)$ be the cumulative distribution function for the data population. Then define $D(x)$

$$D(x) = \sum_{m=-\infty}^{\infty} [F(x10^m) - F(10^m)], \qquad x > 0 \tag{4.24}$$

D(x) is continuous and D(n) for n=2,...,9,10 gives the proportion with leading digit n or less.

Suppose we multiple the data by c, let $D_0(x)$ be the distribution after scaling, similarly, $D_0(x)$ is continuous and $D_0(n)$ for n=2,...,9,10 gives the proportion with leading digit n or less.

So $\sum_{m=-\infty}^{\infty}[F(\frac{n}{c}10^m) - F(\frac{1}{c}10^m)]$ is the proportion whose c-tulping has a leading digit n or less.

$$D_0(n) = \sum_{m=-\infty}^{\infty} [F(\frac{n}{c}10^m) - F(\frac{1}{c}10^m)] \tag{4.25}$$

Assume the distribution does not change, so

$$D(n) = D_0(n) \tag{4.26}$$

$$D(n) = \sum_{m=-\infty}^{\infty}[F(\frac{n}{c}10^m) - F(\frac{1}{c}10^m)] = \sum_{m=-\infty}^{\infty}[F(\frac{n}{c}10^m) - F(10^m)] - \sum_{m=-\infty}^{\infty}[F(\frac{1}{c}10^m) - F(10^m)] \quad (4.27)$$

$$D(n) = D(\frac{n}{c}) - D(\frac{1}{c}) \quad (4.28)$$

According to definition, D(x) is *continuous*, if let $c = \frac{1}{2}$ and $c = \frac{1}{10}$ respectively,

$$D(x) = D(2x) - D(2) \quad (4.29)$$

$$D(x) = D(10x) - D(10) \quad (4.30)$$

and $D(10) = 1$, because all the numbers have a leading digit 9 or less.

Let $H(x) = D(10^x)$, substitute Equation 4.29 and 4.30,

$$H(log\ x) + H(log\ 2) = H(log\ x + log\ 2) \quad (4.31)$$

$$H(log\ x) + H(log\ 10) = H(log\ x + log\ 10) \quad (4.32)$$

Recursively, take Equation 4.31,

$$2H(log\ 2) = H(2log\ 2) \quad (4.33)$$

$$H(2log\ 2) + H(log\ 2) = H(3log\ 2) \Leftrightarrow 3H(log\ 2) = H(3log\ 2) \quad (4.34)$$

....

So we can conclude

$$NH(log\ n) = H(Nlog\ n) \qquad n = 2,\ 10 \quad (4.35)$$

Then take Equation 4.32, and since $H(1) = 1$

$$H(log\ x) + 1 = H(log\ x + 1) \quad (4.36)$$

Let n=10 in Equation 4.35,

$$H(N) = N, \qquad N \in \mathbb{Z} \quad (4.37)$$

As stated in introduction, irrational number can be expressed in a continued fraction form. Hence

$$log\ 2 = \delta_m + o(1/Q_m), \tag{4.38}$$

where $\delta_m = \frac{P_m}{Q_m}$ is the m-th convergent of the continued fraction, and $\lim_{m \longrightarrow \infty} o(1/Q_m) = 0$.

$$log\ 2 = \frac{P_m}{Q_m} + o(1/Q_m) \tag{4.39}$$

$$Q_m log\ 2 = P_m + o(1) \tag{4.40}$$

$$H(Q_m log\ 2) = H(P_m) + o(1) \tag{4.41}$$

$$Q_m H(log\ 2) = P_m + o(1) \tag{4.42}$$

Therefore $H(log\ 2) = log\ 2$.

It has already been proved in Section 4.1.3 that for an irrational number a, $a_n = na - [na]$is uniformly distributed in [0,1]. So for any $0 \leq h < 1$, there exists $a_{n'} = h$. Let $a = log\ 2$,

$$H(a'_n) = H(n' log\ 2 - [n'a]) \tag{4.43}$$

By Equation 4.36,

$$H(t) + 1 = H(t+1) \tag{4.44}$$

$$H(t) + 2 = H(t+1) + 1 = H(t+1+1) = H(t+2) \tag{4.45}$$

$$\ldots$$

$$H(t) + M = H(t+M) \quad\quad M \in \mathbb{Z} \tag{4.46}$$

Hence,

$$H(a'_n) = H(n' log\ 2) - H([n'a]) = n' H(log\ 2) - [n'a] = n' log\ 2 - [n'a] = a'_n. \tag{4.47}$$

Up to now, it has been proven that

$$H(N) = N \quad\quad N \in \mathbb{Z} \tag{4.48}$$

$$H(h) = h \qquad h \in [0,1] \tag{4.49}$$

so for any $y \in \mathbb{R}$, $y = [y] + y - [y]$

$$H(y) = H([y] + y - [y]) = H([y]) + H(y - [y]) = [y] + y - [y] = y \tag{4.50}$$

$$D(10^y) = y \tag{4.51}$$

$$D(y) = log_{10}\ y \tag{4.52}$$

Q.E.D.

Then it is necessary to have a close look at to what degree $log_{10}(y)$ actually approximates to

$$\sum_{m=-\infty}^{\infty} [F(y10^m) - F(10^m)]. \tag{4.53}$$

Let $G(y) = F(10^y)$, $x = log_{10}y$

$$J(x) = \sum_{m=-\infty}^{\infty} [G(x+m) - G(m)] \tag{4.54}$$

Thus, the error will be

$$e = |J(x) - x| \tag{4.55}$$

which is what we want to calculate. Define

$$J(x|t) = \sum_{m=-\infty}^{\infty} [G(x+m) - G(m)]t^{|m|} \qquad 0 < t < 1 \tag{4.56}$$

so that it is convergent and we can apply $Fourier\ Transform$.

$$W(u) = \int_{-\infty}^{\infty} e^{iut} G'(t)dt \tag{4.57}$$

$$G'(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-iut} W(u)du \tag{4.58}$$

$$\int_{-\infty}^{x} G'(t)dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{x} e^{-iut} W(u)dtdu \tag{4.59}$$

$$G(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1-e^{-iux}}{iu} W(u)du \tag{4.60}$$

Hence,

$$J(x|t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1-e^{-iux}}{iu} W(u)P(u,t)du \tag{4.61}$$

where

$$P(u,t) = \frac{1-t^2}{1+t^2-2tcosu}, \qquad \frac{1}{2\pi}\int_{-\pi}^{\pi} P(u,t)du = 1 \tag{4.62}$$

and

$$\lim_{t\to 1} \frac{1}{2\pi}\int_{-\pi}^{\pi} P(u,t)Q(u)du = Q(0) \tag{4.63}$$

Alternatively we can express $J(x|t)$ in the form of

$$J(x|t) = \sum_{k=-\infty}^{\infty} \frac{1}{2\pi}\int_{-\pi}^{\pi} P(u,t)\frac{1-e^{-ix(u+2\pi k)}}{iu+2i\pi k} W(u+2\pi k)du \tag{4.64}$$

$$\lim_{t\to 1^-} J(x|t) = x + \sum_{k\neq 0} \frac{1-e^{-2i\pi kx}}{2i\pi k} W(2\pi k) \tag{4.65}$$

According to $Abel'sTheorem$

$$J(x) = J(x|-1) = x + \sum_{k\neq 0} \frac{1-e^{-2i\pi kx}}{2i\pi k} W(2\pi k) \tag{4.66}$$

$G(x)$ is symmetric about zero, $G(x) = 1 - G(x)$

$$J(x) - x = \sum_{k=1}^{\infty} \frac{sin(2\pi kx)}{\pi k} W(2\pi k) \tag{4.67}$$

Since $|1-e^{-i2\pi kx}| \leq 2$

$$|J(x) - x| \leq \sum_{k\neq 0} \frac{1}{i\pi k} W(2\pi k) \tag{4.68}$$

Previously we define $W(u) = \int_{-\infty}^{\infty} e^{iut} G'(t)dt$

$$\begin{aligned} |W(2\pi k)| &= |\int_{-\infty}^{\infty} e^{i2\pi kt} g(t)dt| &(4.69)\\ &= |\frac{1}{2\pi ki} e^{i2\pi kt} g(t)|_{-\infty}^{\infty} - \frac{1}{2\pi ki}\int_{-\infty}^{\infty} e^{i2\pi kt} g'(t)dt| &(4.70)\\ &= |\frac{1}{2\pi ki}\int_{-\infty}^{\infty} e^{i2\pi kt} g'(t)dt| &(4.71)\\ &\leq \frac{1}{2\pi k}\int_{-\infty}^{\infty} |dg| = \frac{1}{2\pi k} V[g] &(4.72)\end{aligned}$$

Hence,

$$|J(x) - x| \leq \frac{V[g]}{2\pi^2}\sum_{k\neq 0}\frac{1}{k^2} = \frac{1}{6}V[g] \tag{4.73}$$

gives a good approximation since $|J(x) - x|$ is small.

In addition, we take $\frac{1}{2\pi^2}\sum_{k\neq 0}\frac{1}{k^2} = \frac{1}{6}$. Here is the proof.

Consider the function $f(x) = x^2, \ \ -1 \leq x \leq 1$ with period T=2. Represent it in Fourier Series which is

$$x^2 = \frac{1}{3} + \frac{4}{\pi^2}\sum_{n=1}^{\infty}(-1)^n \frac{1}{n^2} cos(n\pi x) \tag{4.74}$$

Let x=1, and we get

$$\frac{1}{6} = \frac{1}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{1}{2\pi^2}\sum_{n\neq 0}\frac{1}{n^2} \tag{4.75}$$

5 Benford's Distribution

In this part, we are going to take a closer look at Benford's Distribution [4]. This concerns not only the frequency of a digit in the first place, but also the second and further.

In Benford's study, he concludes that the frequency of first digits follows a logarithmic relation

$$F_a = \log(\frac{a+1}{a}) \qquad a = 1, 2, \ldots 9, \tag{5.1}$$

from 20229 observations of purely random numbers and formal mathematical tabulations [4]. Now, consider the probability of digit b in the second place if digit a is in the first place. By conditional probability, we have

$$P[b \text{ in the second place}|a \text{ in the first place}] = \frac{P[b \text{ in the second place} \bigcap a \text{ in the first place}]}{P[a \text{ in the first place}]} \tag{5.2}$$

$$b = 0, 1, \ldots 9 \tag{5.3}$$

Note that we will use the same notation for decimal numbers as in Benford's study, which is that a two-digit decimal number with a in the first place and b in the second is written as ab. Therefore, to obtain $P[b \text{ in the second place} \bigcap a \text{ in the first place}]$, we shall consider ab as a whole and follow Benford's logarithm relation of the frequency of the first digit. In this way, we can get the frequency of ab in the first place, which is the equivalent of the frequency of a in the first place and b in the second.

$$P[b \text{ in the second place} \bigcap a \text{ in the first place}] = \log(\frac{ab+1}{ab}) \tag{5.4}$$

Hence, we get the frequency $F_{b|a}$ of second-place digit b following first-place digit a

$$F_{b|a} = \log(\frac{ab+1}{ab})/\log(\frac{a+1}{a}) \tag{5.5}$$

Consider the frequency of digit q in the k-th position following $ab \ldots op$. By using the similar method above, we can get

$$F_{q|ab \ldots op} = \log(\frac{ab \ldots opq+1}{ab \ldots opq})/\log(\frac{ab \ldots op+1}{ab \ldots op}) \tag{5.6}$$

To get the total probability of the digit b in the second place, we shall employ the total probability formula

$$P[b \text{ in the second place}] = \sum_{i=1}^{9} P[b \text{ in the second place } |a_i \text{ in the first place}] \cdot P[a_i \text{ in the first place}] \tag{5.7}$$

$$P[b \text{ in the second place}] = \sum_{i=1}^{9} \log(\frac{a_i b+1}{a_i b}) = \log(\frac{1b+1}{1b} \cdot \frac{2b+1}{2b} \ldots \frac{9b+1}{9b}) \tag{5.8}$$

Table 6: Probability of digit N in the first and second place

N	0	1	2	3	4	5	6	7	8	9
P[first place]	/	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046
P[second place]	0.120	0.114	0.109	0.104	0.100	0.097	0.093	0.090	0.088	0.085

From the table above, we can conclude that the frequency of digit 0 through 9 in the second place is much more uniform than in the first place.

Similarly, the frequency of digit q in a higher (i.e. the k-th) position depends on all the digits in the first through the $k-1$th position. Extend the conclusion we draw about the distribution of the second-place digit, and we can see that when we consider all the possible combinations of $ab \dots op$, the frequency of 0 through 9 in the k-th position is nearly uniform, or

$$F_q = 0.1 \tag{5.9}$$

Therefore, we have come to the conclusion that the frequency of the digits is closely uniform in higher order decimal places because of the effect of taking into account all the possible combinations of the digits that precede it [4].

6 Further Discussions

6.1 Shortcomings of Pinkham's Apporach

To prove Benford's law, Pinkham used the method of scale-invariant. The significance of scale-invariant is that, since god does not choose the unit we are using, the Benford's Law should holds for all unit after re-scale. Since the re-scale simply means the multiply of numbers, a function $D(x)$ is created as following:

$$D(x) = \sum_{m=-\infty}^{\infty} [F(x10^m) - F(10^m)], x > 0 \tag{6.1}$$

where $F(x)$ is the cumilative distribution function for the population of all physical constants.

Then by the scale-invariant:

$$D(n) = D(n/c) - D(1/c), c > 0; n = 2, 3..., 10 \tag{6.2}$$

Then one can deduce $D(n)$=$\log_{10}n$.

However, there are two main shortcomings of Pinkham's theory: nonuniqueness and finite-additivity.

For the nonuniqueness part, Pinkham assume $F(x)$ to be the cumilative distribution function of all physical constants and his prove all depends on there is such $F(x)$. Such $F(x)$ seems to exist only in some God's chamber, we can't know how the graph of $F(x)$ actually looks like but we also can't deny its existence.

However, there are strong evidence indicates that such $F(x)$ doesn't exist. it gives unease. For example, Raimi [5] states that "If h is the real number such that $F(h) = 1/2$, then half the numbers in the universe are less than h, which makes h a most remarkable physical constant. Now what becomes of h when we exxercise our freedom to make scale change?". D.Knuth [6] also points out that the mathematics alone can prove no such F exists. Using b to represent the base for our numeration system, Pinkham's prove shows that:

$$\sum_{m=-\infty}^{\infty}[F(rb^n) - F(b^n)] = log_b r \tag{6.3}$$

for all integers r and b with $1 < r < b$. It easy to prove no such F exist.

Pinkham states that The logarithmic "law" shows $D(n)$ should be approximately $\log_{10}(n)$, then using scale-invarlant, the second formula in this text. However, Pinkham [2] assume $D(n)$ to be continuous but using $n = 2, 3, 4..., 10$ to deduce $D(n)$ =$\log_{10}(n)$. It's clear that since n is only a set of discrete integars point, there are infinte $D(n)$ which satisfy $n = 2, 3..., 10$, as is shown in Figure 3.

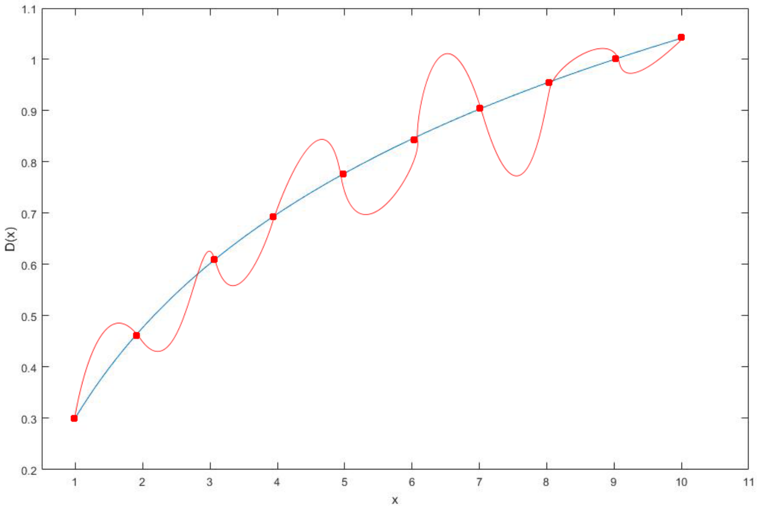

Figure 3: Several examples of $D(n)$ that satisfy the condition

Therefore we can conclude that $D(n)$ is not unique, which is a shortcoming of Pinkham's prove.

For the finitely addtive part, Pinkham [2] assumes $F(x)$ to be population of all physical constants. But all these constants are necessarily only finitely additive, thus lack of countable additivity. The constants themselves are not well defined to culmulate the density of distribution, since the underlying set is countable and "the density of each singleton is 0" [7]. For example , consider the set of positive integers F_d with first significant digit being d. But since the upper limit is not equal to the lower limit, thus the function of nature density does not converge, thus doesn't exist. For example, as $d = 1$, the upper limit is $\frac{5}{9}$ while the lower limit is $\frac{1}{9}$ [5].

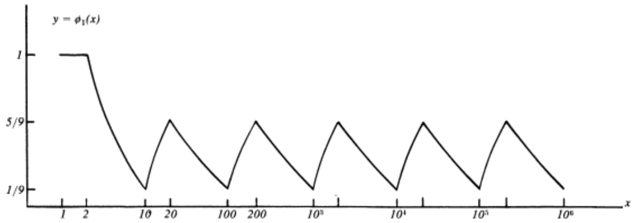

Figure 4: Density for F_d $(d = 1)$

6.2 Hill's proof

The main idea of Hill's [7] prove is "base-invariant" and an appropriate measurability structure, as well as the definition below, which make up the shortcomings of nonuniqueness and finite-additivity of the former proof. The first process is to define the mantissa σ-algebra below:

$$M_b(x) = r, r \in [1, b), x = rb^n \tag{6.4}$$

The (base b) mantissa σ-algebra, $\mathscr{M}_b$ is the σ-algebra on $\mathbf{R}^+$ generated by M_b. For example, $M_{10}(9) = 9 = M_{100}(9), M_2(9) = 9/8$.

Thus there is an important lemma which "will be a key ingredient in the definition of base-invariance below" [7]

$$< E >_b = \cup_{k=0}^{n-1} < b^k E >_{b^n} \tag{6.5}$$

for all $n \in \mathbf{N}$ and $E \subset [1, b)$.

Thus it avoid the problem of the existence of $F(x)$, since finite intervals like (c, d) are not in $\mathscr{M}_b$, so we will not consider such universal median h of $F(h) = 1/2$. In order to generate $\mathscr{M}_b, D_b^{(i)}(x)$ is defined to represent the ith significant dight of x (base b). A mathematic way formally define $D(x)$, and $D(x)$ only care about the significant dight, ignoring the decimal point to aviod the shortcoming of finitely additive. Also it is continous, since all real number including irrational number can be presented in such way. For example, $D_1^{(1)}0(\pi) = 3, D_2^{(2)}(\pi) = 1$.

According to the generalized significant digit law [5], then we can get that for every integer $b > 1$,

$$P(\bigcap_{i=1}^{k} D_b^{(i)} = d_i) = log_b[1 + (\sum_{i=1}^{k} b^{k-i} d_i)^{-1}] \tag{6.6}$$

Thus, since $\mathscr{M}_b$ is the σ-algebra generated by $D(x)$. The generalized significant digit rule suggests it holds for all d_i where $i = 1, 2, 3...k$, the former formula is just the special case of this law. "any probability measure on the measurable space($\mathbf{R}^+$, $\mathscr{M}_{10}$) will uniquely determine the probability of such an event" [7] , which prove the uniqueness.

To get the Borel probability measure $\hat{P}$ on $[1, b)$, Hill uses a similar method as the "disk distribution" metioned in Pinkham's proof [2].

$$P(< E >_b) = \hat{P}(E), E \in \mathbf{B}[1, b) \tag{6.7}$$

It is easy to see that there is no scale-invariant(countably-additive) Borel probability measure on the positive reals [5]. Hill [7] uses base-invariance to deduce the benford's law. Similar to the disk distribution, this relationship defines a "1-1 correspondence" from P to $\hat{P}$, which is the borel probability measures[7]. By Lemma 2.3 [7]: $\mathscr{M}_b \subset \mathscr{M}_b^n \subset \mathbf{B}$ for all n $\in \mathbf{N}$.

So the probability measure P on ($\mathbf{R}^+$, $\mathscr{M}_{10}$) is $base-invariant$ if $\hat{P}$ satisfies $\hat{P}[1, b^a) = \sum_{k=0}^{n-1} \hat{P}[b^{k/n}, b^{(k+a)/n})$ for all $n \in \mathbf{N}$ and all $a \in (0, 1)$. It's very important since we want to show the $base - invariant$, and this formula offer a relationship between base b and b^n.

Meanwhile using another "P" be the probability measure defined by

$$P_b(< [1, \gamma) >_b) = log_b\gamma \tag{6.8}$$

P_b satisfy Equation 6.6 and precisely defined by the generalized significant digit law. It is also the unique scale-invariant probability measure on ($\mathbf{R}^+$, $\mathscr{M}_b$) . Let $\overline{P}$ be the b-logarithmic rescaling of P, so we have the following formula:

$$\overline{P}[o, a) = \hat{P}[1, b^a) = P(< [1, b^a) >_b) \tag{6.9}$$

Thus by scale-invariant we can deduce it's base-invariant. Then recall the "mod 1" method, "A borel probability measure $\overline{P}$ is invariant under the mapping $nx(\text{mod } 1)$ for all $n \in \mathbf{N}$ if and only if

$$\overline{P} = q\delta_0 + (1 - q)\lambda \tag{6.10}$$

for some $q \in [0, 1]$ " [7].

Let P_* be the probability measure which equals 1 when 1$\in$ E and 0 otherwise. By showing P_* is base-invariant we then discuss the condition for P when it is base-invariant

Here, if $P = qP_* + (1 - q)P_b$, then by the base-invariance of both P_* and P_b, we can deduce that P is base-invariant. The use of the mantissa σ-algebra enable us to use clear countable additive here, rather than finite additive. Then it implies the correctness of the main theorem. Therefore, we can conclude that the Hill's proof use base-invariant to include both discrete and continous data, meanwhile to aviod the shortcomings of nonuniqueness and finite-additivity, and to offer a "full" proof of Benford's Law.

References

[1] 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.

[2] Roger S.Pinkham, On the distribution of first significant digits, *Ann. Math. Statist.*, 32(4):12231230, 12 1961. `http://dx.doi.org/10.1214/aoms/1177704862.`

[3] Hazewinkel, Michiel, ed. (2001), "Continued fraction", *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4

[4] Frank Benford. The law of anomalous numbers. *Proc. Amer. Math. Soc., 123:887895, 1995.* `http://www.jstor.org/stable/984802.`

[5] R. Raimi, The first digit problem, *Amer. Math.* Monthly 83 (1976), 521-538 `http://www.jstor.org/stable/2319349`

[6] Donald. Knuth, The Art of Computer Programming, Vol. 2, Addison- Wesley, Reading, *Mass.*, 1969. pp. 219-229.

[7] Theodore P. Hill. Base-invariance implies Benfords *law. Proc. Amer. Math. Soc.*, 123:887–895, 1995.`http://www.ams.org/journals/proc/1995-123-03/S0002-9939-1995-1233974-8/.`