Distribution Agreement

Mengqi Zhao April 14th, 2015

Study of Benford's Law

By

Mengqi Zhao

Ken Ono

Advisor

Department of Mathematics and Computer Science

Ken Ono

Advisor

Bree Ettinger

Committee Member

David Jacho-Chavez

Committee Member

2015

Study of Benford's Law

By

Mengqi Zhao

Ken Ono

Advisor

An abstract of

a thesis submitted to the Faculty of Emory College of Arts and Sciences

of Emory University in partial fulfillment

of the requirements of the degree of

Bachelor of Sciences with Honors

Department of Mathematics and Computer Science

2015

Abstract

Study of Benford's Law

By Mengqi Zhao

Having been studied for over a hundred years, Benford's Law is an anomaly of numbers. It was named after Frank Benford and was introduced in the 19th century. In this paper, we illustrate its counter-intuitive nature that numbers are actually not distributed with equal probability in Section 1. The law started with the study of the leading digit frequency, but expanded to the second digit, the first-two digits, the first-three digits, . . . , and to any base beyond base 10. Though the real-world applications of the law were limited before the 20th century, it has been used by auditors, accountants, scientists to detect data fraud in recent times. This will be discussed in Section 1 following the background information.

In Section 2, by introducing the concept of uniform distribution mod 1, we will define the notion of a Benford sequence. The next step is to establish the mathematical foundations by defining a good sequence and Weyl's Criterion in the mathematical justification subsection. They provide a strategy to prove whether a sequence conforms to Benford's Law or not. Then the strategy was proved accordingly.

Finally, with the established strategy, we take a look at how mathematicians have used it to prove some sequences to be Benford.

Study of Benford's Law

By

Mengqi Zhao

Ken Ono

Advisor

A thesis submitted to the Faculty of Emory College of Arts and Sciences

of Emory University in partial fulfillment

of the requirements of the degree of

Bachelor of Sciences with Honors

Department of Mathematics and Computer Science

2015

Contents

List of Figures

List of Tables

1 Introduction

Big data problems are often difficult to analyze and solve. Statisticians often study large data sets with millions of data pieces. They try to identify distributions by calculating various statistics including mean, variance, standard deviation, skewness, etc. From the distributions of various data resources, statisticians are able to answer a lot of questions related to the given data. They can build corresponding models, analyze human behaviors underlying the experimental data, forecast market performances from stock prices, or solve other important problems concerned with certain data.

For a very long time, there were very few peculiar data patterns found by data scientists. Examples could be drawn easily. Since any integer is either even or odd, we expect that a big database with only integers would contain equal amount of even and odd numbers. In "real-world" large databases, even numbers and odd ones are indeed roughly equally-distributed just as expected. Table 1 below gives a list of 60 countries sorted by 2013 nominal GDP in millions of US dollars by the United Nations. After doing a simple count, we found 26 even numbers, which is nearly one half of the total number! Moreover, a very big database theoretically follows the Central Limit Theorem (CLT), stating that a random sample with sufficiently many observations will be approximately normally distributed. Not only statisticians, but also economists and experts from other fields have utilized this theorem in real-life data analysis. It is a fact that many statistical tests in economics have been built with the assumption of the CLT.

However, many natural data sets exhibit unusual behavior. Unlike even and odd numbers, the significant digits which usually describe the size of a number, for example the first digit, were found to follow a counter-intuitive pattern. One might naively expect each digit from 0 to 9 to occur with equal frequency. This is also saying that, in a big database, one-ninth of the numbers should start with 1 or 2, etc. Each digit from 0-9 is also expected to have a probability of one-tenth to be on the second digit position. However, it turns out to be a completely different story.

Country/Region	GDP(Millions of US$)	Country/Region	GDP(Millions of US$)
United States	16,768,100	United Arab Emirates	402,340
China	9,181,204	Colombia	378,148
Japan	4,898,532	Venezuela	371,339
Germany	3,730,261	South Africa	366,060
France	2,806,432	Denmark	336,701
United Kingdom	2,678,455	Malaysia	312,434
Brazil	2,243,854	Singapore	295,744
Italy	2,149,485	Israel	291,567
Russia	2,096,774	Chile	277,043
India	1,937,797	Hong Kong	274,027
Canada	1,838,964	Philippines	272,067
Australia	1,531,282	Finland	267,329
Spain	1,358,263	Egypt	255,199
South Korea	1,304,554	Greece	241,721
Mexico	1,259,201	Ireland	232,077
Indonesia	868,346	Portugal	227,324
Netherlands	853,539	Pakistan	225,419
Turkey	822,149	Kazakhstan	224,415
Saudi Arabia	748,450	Czech Republic	208,796
Switzerland	685,434	Algeria	208,764
Argentina	611,727	Qatar	202,450
Sweden	579,680	Peru	200,269
Poland	525,863	Irap	195,517
Belgium	524,806	New Zealand	189,025
Norway	522,349	Romania	188,881
Nigeria	514,965	Ukraine	182,026
Iran	492,783	Kuwait	175,831
Taiwan	489,089	Vietnam	171,222
Austria	428,322	Bangladesh	153,505
Thailand	420,167	Hungary	129,989

Table 1: List of Countries with Top 60 GDP by the United Nations (2013) [4]

Ordinary statistical analysis of a given data set like above may include answering such questions: What is the average national GDP level among the top 60 countries? A number is randomly drawn from the table, what is probability that it is an even number? How large is the gap between some wealthier and poorer countries? What does the distribution of national GDP look like and at what extent does GDP vary from country to country? Many of the common statistical questions could be answered by calculating the statistics.

$$\text{Mean} = 1,205,468$$
$$\text{Range} = 16,638,111$$
$$\text{Standard Deviation} = 2496543.821$$
$$\text{Skewness} = 4.87144.$$

Average national GDP of the 60 countries in Table 1 is $1,205,468$ and the GDP gap between the 1st ranked country (the US) and the 60th ranked country (Hungary) is 16,638,111. A high standard deviation implies that the GDPs of the first 60 countries spread out over a wide range. In addition, a positive skewness indicates that the mass of the data set distribution is concentrated on the left. The probability distribution function has a longer or fatter right tail than the left one.

Unlike the ordinary pattern and the traditional statistical analysis, the distribution of significant digits is truly odd. We will take the leading digit as an example. Intuitively, there are 9 candidates (number 1-9) for the first digit of a number. According to naive intuition, the chance for a number to start with 1 should be one-ninth. However, Table 2 and Figure 1 below tell us that the digit distribution of the GDP data is not as uniform as expected. For example, 25% of the 2013 national GDP numbers start with 1, which is more than twice of the expected 1/9. Moreover, 35% of the 60 numbers in Table 1 start with 2, but only about 1.7% start with 9. In the full list of 194 countries on the United Nations' website, 56 numbers (28.8%) have the first digit as 1. Instead of an equal distribution, lower digits (1 and 2) appear more often as first digits than high digits (8 and 9). This counter-intuitive phenomenon is known as Benford's Law.

Digit	1	2	3	4	5	6	7	8	9
Frequency	15	21	6	6	5	2	1	3	1
Proportion	0.25	0.35	0.1	0.1	0.083	0.033	0.0167	0.05	0.0167

Table 2: Digit Frequency Table of the Data in Table 1

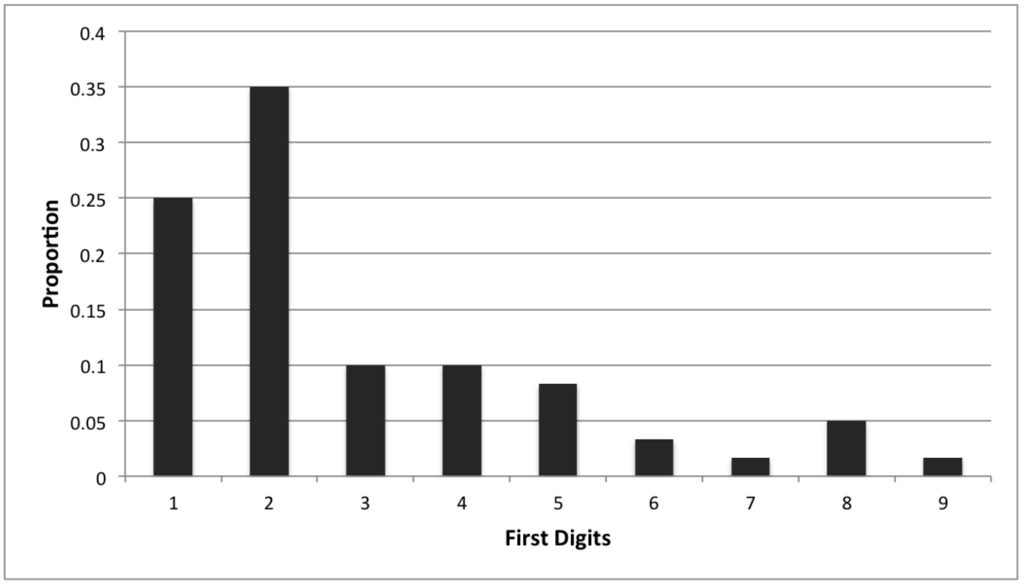

Figure 1: First Digit Graph of the Data in Table 1

1.1 Historical Background

The establishment of Benford's Law dates back to 1881. This statistical principle was first proposed by Simon Newcomb, who was well-known as a highly-honored Canadian-American astronomer. Newcomb made a significant contribution to the derivation of physical constants and planetary motion in early 19$^{\text{th}}$ century. However, he focused more on mathematics later on in his life [6]. He first noticed a peculiar phenomenon while studying logarithm tables that contain a logarithm of numbers starting with lower numbers. Newcomb found that the earlier pages of the tables were much more worn than later ones. Apparently, people more often referred to earlier pages, which contain numbers beginning with lower digits such as 1 or 2. He published his observation in the American Journal of Mathematics, where he proposed that the probability of a number N being the first digit of a number was: Prob($D_1 = N$)=$\log_{10}(N+1) - \log_{10}(N)$. This means that in base 10, a number that starts with 1 would appear approximately 30% of the time. A number that starts with 2 would then appear around 17.6% of the time. The frequency monotonically decreases as N increases. Newcomb did not show the equation explicitly, but he was obviously aware of it because of the probability table he included in the paper [8] :

N	Prob($D_1 = N$)	Prob($D_2 = N$)
0	-	0.1197
1	0.3010	0.1139
2	0.1761	0.1088
3	0.1249	0.1043
4	0.0969	0.1003
5	0.0792	0.0967
6	0.0669	0.0934
7	0.0580	0.0904
8	0.0512	0.0876
9	0.0458	0.0850

Table 3: Expected Digits Frequency Derived By Newcomb

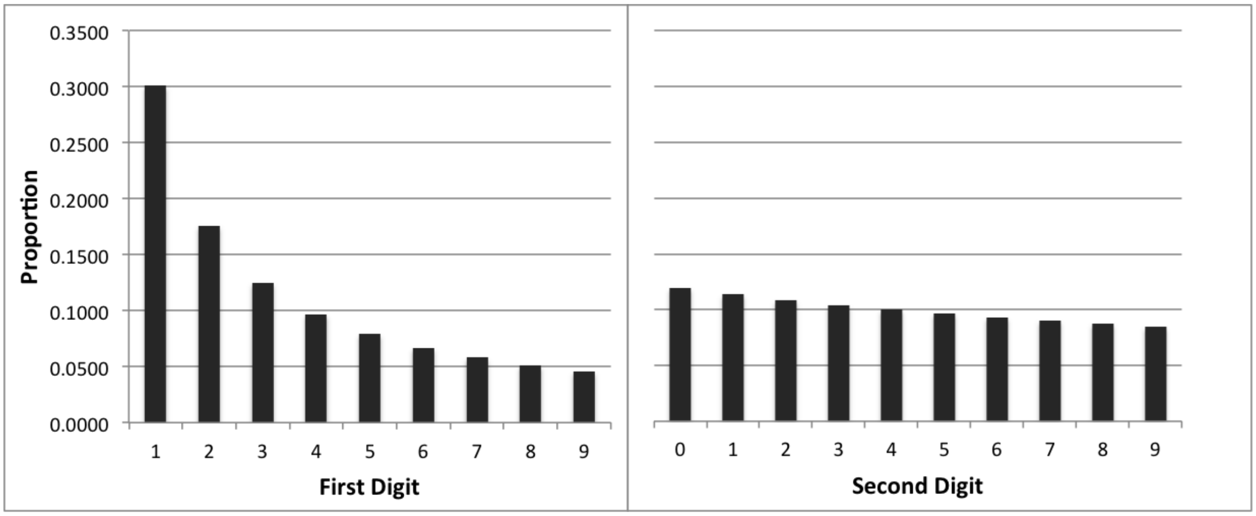

Figure 2: Expected First and Second Digits Distribution Derived by Newcomb

Completely deviating from the intuition: Prob($D_i = N$)= 1/9, for $N = 0, 1, \ldots, 9, i \in \mathbb{N}$, except Prob($D_1 = 0$) that is not applicable, Newcomb's observation was just as astounding as what we found in the previous GDP table. However, since he didn't explain the phenomena theoretically, his article did not draw much attention at the time. Newcomb would be surprised to see how extensively it has been studied since then [6].

Later in 1938, Frank Benford again mentioned the same situation about the logarithm tables in his *The Law of Anomalous Numbers* paper [2]. As an illuminating engineer and physicist, Benford had 20 patents and published more than 100 papers on lights and the science of optics. His study on digits originated from his interest in mathematics as a hobby. His patents expired a long long time ago, however, Benford's study of digit distribution has lived on and made a great impact on subsequent related research. Benford's first step was to analyze the first digit frequency in 20 data tables from various natural sources. The results of his analysis are displayed in the following table (Table 2). He confirmed that more numbers tend to begin with lower first digits (1 and 2) instead of higher digits (8 and 9). Beyond that, the expected frequencies of the digits in number lists including the first digit, the second digit, the first-two digits, and so on were also derived [9]. Frank Benford's re-discovery of the principle, his mathematical analysis, and his experiment on 20229 natural numbers and 20 data sets made the law so well-known that it was given his name.

First Digit (in Percentage)											
Group	**Description**	**Count**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
A	Rivers, Area	335	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1
B	Population	3,259	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2
C	Constants	104	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6
D	Newspapers	100	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0
E	Spec. Heat	1,389	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1
F	Pressure	703	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7
G	H.P.Lost	690	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6
H	Mol.Wgt.	1,800	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2
I	Drainage	159	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9
J	Atomic Wgt.	91	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5
K	$n^{-1}, \sqrt{n}, \ldots$	5,000	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9
L	Design	560	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6
M	Digest	308	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2
N	Cost Data	741	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1
O	X-Ray Volts	707	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8
P	Am. League	1,458	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0
Q	Black Body	1,165	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4
R	Addresses	312	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0
S	$n1, n2 \ldots n!$	900	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5
T	Death Rate	418	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1
	Average	1,011	30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7
Probable Error			±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.3

Table 4: Benford's Analysis of Natural Data Sets in 1938 [2]

1.2 Real-life Applications

Benford's Law was not the center of the public's attention for a long time because of its lack of applications. However, this is not the case anymore. In"The Running Man" episode of a popular television crime drama *NUMB3RS*, one character called Charlir Eppes used Benford's Law to solve a series of crimes. This is just an interesting use of Benford's Law on television. In real life, major fields of the law's applications include: accounting, auditing, and detection of anomalies in data sets.

Not until the 1980s had Benford's Law been utilized to test the validity of natural data resources. Two remarkable digital analyses on income statements came from Carslaw and Thomas at that time. Carslaw, in 1988, detected abnormal income manipulation behavior from a New Zealand firms' financial statements. Their stated earnings had more zeros as the second digit than expected by the Benford frequency, but at the same time, there were too few nines. As we normally perceive a price of $1.99 to be much cheaper than $ 2.00, companies also have the tendency to round up their earnings to make better-looking statements [3]. From the observed second digit frequencies, Carslaw argued that those firms rounded up earnings such as $ 1,900,000 to $2,000,000 to enhance income numbers. Even though he utilized Benford's expected digit frequency, Carslaw did not refer it to Benford's Law, but to "Feller's Proof". One year later, Thomas studied the U.S. firms' earnings and noticed a similar pattern as what Carslaw found. The first person who utilized Benford's Law substantially as a key indicator of accounting fraud was Mark J. Nigrini, an accounting professor at West Virginia University. His assertion about the important role of Benford's Law in auditing started from 1996, when he identified tax evaders. He and the co-author, Mittermaier, further explained the process of auditing accounting data by Benfod analysis in details and illustrated some case studies in their 1997 paper. Nigrini also made the law and its use popular through his recent book *Benford's Law: Applications for Forensic Accounting, Auditing, and Detection of Data Frauds* [5].

Empirically, people might tend to manipulate the distribution of the first digits in financial accounts to be more uniform if they want to fraud. Based upon that, auditors could perform a digital analysis on the target accounts to distinguish any abnormal digits distribution. In the paper *The Effective Use of Benford's Law to Assist in Detecting Fraud in Accounting Data*, Cindy Durtschi and her colleagues conducted an analysis on two financial accounts of a large Western U.S. medical center. From the results of their analysis of the office supplies reimbursement account, they found that except digits 2 and 7, the other seven digits almost conformed to the expected Benford distribution. Their further study on non-conforming items went against the existence of accounting frauds; The payments were proved legitimate. However, the insurance refund check account's number deviations were not as normal as the office supplies refunds.

The overall insurance refunds deviations did not fall within the conforming ranges except for the digit 2. Durtschi and her colleagues found that in the entire account many more checks were written at just over $ 1,000 than in the previous periods and than the expected Benford frequency. Previously, most checks were actually less than $100.00 . Although the financial officer claimed that she tried to write fewer checks and refund in an accumulated amount to some large insurers, it turned out that she owned a shell company and had written excessive refund checks to her own shell company. This example illustrates that Benford's Law is very useful for detecting suspicious account activities. At the same time, auditors and accountants should pay attention to the adaptability of the law to different types of accounts. For example, Benford testing may work well with income, expenses, disbursements, transactions numbers, etc. However, it is not likely for Benford analysis to be useful for the data sets with pre-assigned numbers, such as ATM withdrawals. Patient refund accounts of a medical center are also unlikely to conform since co-payments were usually assigned through insurance plans in advance. Moreover, accounts that only record numbers that fall into a certain range, such as accounts with a built-in maximum or minimum, would not be in the scope of Benford analysis [5].

Beyond spotting fraudulent accounting data and financial statements, earth scientists and other users of earth science data also would like to verify the validity and accuracy of the data they found. Here comes Nigrini's study on U.S. streamflows to show whether the data sets conform to Benford's Law or not. The National Streamflow Information Program (NSIP) of the U.S. Geological Survey (USGS) collects data of water flows at stream gauge sites over major floods and droughts. The flow data is important because they could help other government departments respond to the disasters appropriately and effectively. Moreover, the needs from designers to build bridges and other facilities enhance the significance of the streamflows data; As do their usefulness to assess the survivability of endangered animals during extreme conditions, to forecast future flows, and to monitor water quality. As a result, testing the credibility of the given earth science data before making major decisions from the data becomes critical. Nigrini performed a first-two digits frequency analysis on 457,440 usable stream gauge measurements, which covered an extended period of 130 years. It was found that the actual frequency conformed nearly perfectly to the expected frequency. This indicates that earth science data should follow Benford's Law, which is useful to examine the authenticity of the data. If close conformity is observed, accuracy and integrity of the data can be confirmed and the data can be forwarded to further uses [9].

Having introduced the origins of Benford's Law and its wide-ranging applications in the real life, we can move on to study its mathematical aspect and substantiate its mathematical foundations.

2 Mathematical Foundations of Benford's Law

To properly define Benford's Law and perform a study on its mathematical justification, it turns out that we need the concepts of uniform distribution mod 1. So in this section, we will first familiarize ourselves with uniform distribution mod 1 and bases in mathematics, which will build the foundations. Then, we will be ready to define Benford's Law in base B. How uniform distribution establishes and justifies Benford's Law will be illustrated. In simple terms, if the mantissas of the logarithm of a sequence are uniformly distributed, the sequence will follow the Benford distribution. Based upon this idea, we will visit Weyl's Criterion, which provides us with a strategy to prove uniform distribution modulo 1. Thus, we would be able to further determine whether Benford's Law holds for a sequence or a class of sequence.

2.1 Uniform Distribution Modulo 1

Let $[x]$ and $\{x\} = x - [x]$ be defined as the *integral part* and *fractional part* of $x \in \mathbb{R}$, respectively. Note that $\{x\}$ is also called mantissa of x or the residue of x (mod 1), $\{x\} \in$ *unit interval* $I = [0, 1)$.

Let $\omega = (x_n), n = 1, 2, \ldots,$ be a given sequence of real numbers. For a positive integer N and a subinterval $[a, b) \subseteq I$, we define a counting function $A([a, b); N; \omega)$ as the number of terms $x_n, 1 \leq n \leq N$, for which $\{x_n\} \in [a, b)$.

Definition 1 (Definition 1.1 in [7]). *The sequence $\omega = (x_n), n = 1, 2, \ldots,$ of real numbers is defined to be **uniformly distributed modulo 1** (abbreviated u.d. mod 1) if for every pair $a, b \in \mathbb{R}$ with $0 \leq a < b \leq 1$ we have*

$$\lim_{N \to \infty} \frac{A([a, b); N; \omega)}{N} = b - a. \tag{2.1}$$

This is also saying that if each subinterval $[a, b) \subseteq I$ obtains its deserved share of the fractional parts, the sequence $\omega = (x_n), n = 1, 2, \ldots,$ is u.d. mod 1.

To be more general, let $c_{[a,b)}$ be the characteristic function of the interval $[a, b) \subseteq I$. Then the equation (2.1) could be modified to the following form:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} c_{[a,b)}(\{x_n\}) = \int_0^1 c_{[a,b)}(x)dx \tag{2.2}$$

This follows by the theorem below.

Theorem 1 ([Theorem 1.1 in [7]). *The sequence $(x_n), n = 1, 2, \ldots,$ of real numbers is u.d. mod 1 if and only if for every real-valued continuous function f defined on the closed unit interval $\bar{I} = [0, 1]$ we have*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(\{x_n\}) = \int_0^1 f(x)dx \tag{2.3}$$

Proof. First, let (x_n) be u.d. mod 1 and $f(x) = \sum_{i=0}^{k-1} d_i c_{[a_i,a_{i+1}]}(x)$ be a step function on $\bar{I} = [0, 1]$, where $0 = a_0 < a_1 < \cdots < a_k = 1$. Then (2.2) indicates that for every such f equation (2.3) holds.

We assume that f is a real-valued and continuous function on $\bar{I}$. The Riemann integral defines $\int_a^b f(x)dx$ to be the area under the function graph over the interval $[a, b]$ based upon partition of $[a, b] : P = \{a = u_0 < u_1 < \cdots < u_k = b\}$. A function f on $[a, b]$ is a step function if there exists a partition P. It follows that given any $\epsilon > 0$, there exist two step functions, f_1 and f_2 such that $f_1(x) \leq f(x) \leq f_2(x)$ for all $x \in \bar{I}$ and $\int_0^1 (f_2(x) - f_1(x))dx \leq \epsilon$. We would derive inequalities as following:

$$\int_0^1 (f(x) - f_1(x))dx \leq \epsilon$$

$$\int_0^1 f(x)dx - \epsilon \leq \int_0^1 f_1(x)dx = \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} f_1(\{x_n\})$$

$$\leq \underline{\lim_{N\to\infty}} \frac{1}{N}\sum_{n=1}^{N} f(\{x_n\}) \leq \overline{\lim_{N\to\infty}} \frac{1}{N}\sum_{n=1}^{N} f(\{x_n\})$$

$$\leq \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} f_2(\{x_n\}) = \int_0^1 f_2(x)dx \leq \int_0^1 f(x)dx + \epsilon.$$

Since ϵ is arbitrarily small, (2.3) holds for a real-valued continuous function f.

Conversely, for a sequence x_n, we first assume that (2.3) holds for every real-valued continuous function f defined on $\bar{I}$, and $[a, b) \subseteq I$. For any $\epsilon > 0$, there exist two continuous functions, g_1 and g_2 that $g_1(x) \leq c_{[a,b)}(x) \leq g_2(x)$ for $x \in \bar{I}$. Then we have a chain of inequalities again:

$$\int_0^1 (g_2(x) - g_1(x))dx \leq \epsilon$$

$$b - a - \epsilon \leq \int_0^1 g_2(x)dx - \epsilon \leq \int_0^1 g_1(x)dx = \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} g_1(\{x_n\})$$

$$\leq \underline{\lim_{N\to\infty}} \frac{A([a,b);N)}{N} \leq \overline{\lim_{N\to\infty}} \frac{A([a,b);N)}{N} \leq \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} g_2(\{x_n\})$$

$$= \int_0^1 g_2(x)dx \leq \int_0^1 g_1(x)dx + \epsilon \leq b - a + \epsilon.$$

Since ϵ is arbitrarily small, we here proved (2.3). ■

Corollary 1.1 (Corollary 1.2 in [7]). *The sequence (x_n) is u.d. mod 1 if and only if for every complex-valued continuous function f on $\mathbb{R}$ with period 1 we have*

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} f(x_n) = \int_0^1 f(x_n)dx. \qquad (2.4)$$

Proof. Apply Theorem 1 to the real and imaginary part of function f, we first show that it also holds for complex-valued f. Since here function f has the period 1, then $f(\{x_n\}) =$

$f(x_n)$, which leads to (2.4). In the proof of Theorem 1, we could choose functions g_1 and g_2 such that $g_1(0) = g_1(1)$ and $g_2(0) = g_2(1)$, so that (2.4) can be applied to the periodic extension of g_1 and g_2 to $\mathbb{R}$. We prove the sufficiency [7]. ■

2.2 Definition of "Benford's Law Base B"

In the previous introduction section, stories and examples were given in base 10, which is the most common base used today. Now, we are going to expand to other bases and define Benford's Law in any base.

Definition 2. *A **base** is the number of different digits that a numeral system has to represent numbers. Let B denote a specific base, $B \in \mathbb{N}^* = \{1, 2, \dots\}$.*

For example, base 10 refers to the decimal system that we use most of the time in real life. The decimal system has 10 digits, which are 0,1, 2, . . . , 9. Different arrangements and combinations of these 10 digits can represent different numbers. Unlike human beings, computers use completely different bases such as base 2 under the binary system, with which people who have learned computer science might be familiar. Due to their information processing ability, computers are only able to deal with 0s and 1s. Quinary system (base 5) would be another numerical system, using digits 0 to 4. Note that we will denote numbers with their bases as following: $13_{10} = 1101_2 = 23_5$ and after. In general, a number in base B could be calculated in the following way: $x_1 B^j + x_2 B^{j-1} + \cdots + x_k B^{j-(k-1)}$ for some integer j, $x_k = 1, 2, \dots, B - 1$. We have made a table to show first 12 distinct numbers of the Fibonacci sequence in decimal, binary, and quinary systems for you to review. It is always more straightforward to understand bases and how each system represent numbers differently by a direct comparison.

Decimal(base-10)	**Binary(base-2)**	**Quinary(base-5)**
1	1	1
2	10	2
3	11	3
5	101	10
8	1000	13
13	1101	23
21	10101	41
34	100010	114
55	110111	210
89	1011001	324
144	10010000	1034
233	11101001	1413

Table 5: First 12 Distinct Numbers in Fibonacci sequence with Different Bases 10,2,5

Here follows the definition of Benford:

Definition 3. *For a sequence $\omega = (x_n), n = 1, 2, \ldots,$ and $x_n \in \mathbb{R}$, let*

$$B(d, N, B; \omega) = \frac{\#\{n \leq N : \textit{first digits of } x_n \textit{ in base } B \textit{ are the string } d\}}{N}.$$

*For all $B \geq 2$, the sequence $\omega = (x_n), n = 1, 2, \ldots, x_n \in \mathbb{R}$ is **Benford**, if*

$$\lim_{N \to \infty} B(d, N, B; \omega) \equiv \log_B(d+1) - \log_B(d) \textit{(mod 1)}.$$

Study of leading digits only considers the positive numbers without losing generality. The above equation must hold for any initial string d in any base B for a sequence or function to be Benford. For instance, the function $B(100, N, 2; \omega)$ calculate the proportion of $n \leq N$, for which x_n starts with the string d = "100" in base 2. Since $100_2 = 4_{10}, (\log_2(5) - \log_2(4))$ (mod 1) $\approx$ 0.3219, 32.19 % of a Benford sequence would be expected to start with "100 ". We could also derive the expected digit frequencies of the second digit by following Frank Benford's research [1].

Let D_1, D_2, D_1D_2 denote as the first digit, the second digit, and the first-two digits of a number, respectively.

$$\text{Prob}(D_1 = N_1) = \log_B(N_1 + 1) - \log(N_1);$$

$$\text{Prob}(D_2 = N_2) = \sum_{D_1} \log_B(N_2 + 1) - \log_B(N_2);$$

$$\text{Prob}(D_1 D_2 = N_1 N_2) = \log_B(N_1 N_2 + 1) - \log_B(N_1 N_2).$$

2.3 Mathematical Justification of Benford's Law

Here we are going to show how essential and necessary the uniform distribution mod 1 is for our definition of Benford to hold.

Definition 4. *An integer-valued function x_n is **good** whenever*

$$x_n \sim a_n e^{b_n},$$

which means that

$$\lim_{n\to\infty} \frac{x_n}{a_n e^{b_n}} = 1$$

and the following conditions are satisfied:

(1) There exists some integer $h \geq 1$ such that b_n is h-differentiable, ${b_n}^{(h)}$ is monotone, and

$$\lim_{n\to\infty} {b_n}^{(h)} = 0. \tag{2.5}$$

(2)

$$\lim_{n\to\infty} n\left|{b_n}^{(h)}\right| = \infty. \tag{2.6}$$

(3)

$$\lim_{n\to\infty} \frac{D^{(h)} \log a_n}{{b_n}^{(h)}} = 0, \text{ where } D^{(h)} \text{ denotes the } h^{th} \text{ derivative.} \tag{2.7}$$

Theorem 2. *If x_n is good, then the sequence (x_n) is Benford.*

In the article *The Distribution of Leading Digits and Uniform Distribution Mod 1*, Diaconis concluded that a real sequence (x_n) is Benford, if and only if $\log_B(x_n)$ is uniformly distributed mod 1 for all B. Here we illustrate a numerical example in base 10 to justify Diaconis' result. Table 6 shows logarithms of the 9 possible first digits in base 10. Then we map it onto a numerical axis as Figure 3.

Digits N	$\log_{10}(N)$
1	0
2	0.301
3	0.477
4	0.602
5	0.698
6	0.778
7	0.845
8	0.903
9	0.954
10	1

Table 6: Logarithms of Numbers 1-10 in Base 10

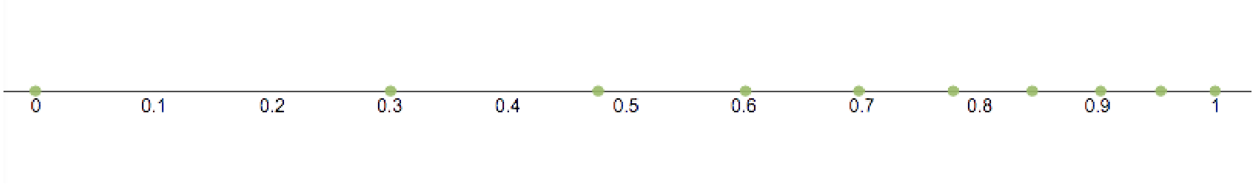

Figure 3: Logarithm Scale Bar

The mathematical basis of Benford's Law is that the mantissas of logarithms of the numbers are uniformly distributed. In base 10, it is also saying that any subinterval $[a, b) \subseteq [0, 1)$ gets its proper share.

Suppose a number $M = 284.67_{10}$. Then, we have

$$\log_{10}(M) = \log_{10}(284.67) = \log_{10}(2.8467 \times 10^2) = 2 + \log_{10}(2.8467)$$

Thus the mantissa of $\log_{10} 284.67$ equals to the mantissa of $\log_{10} 2.8467$. In like manner, $\{\log_{10} 1\} = \{\log_{10} 10\} = \{\log_{10} 100\} = 0, \{\log_{10} 2\} = \{\log_{10} 20\} = \{\log_{10} 200\} \approx 0.301$, etc. Since the logarithm function is a monotonically increasing function, numbers with mantissas less than $\{\log_{10} 2\}$ should start with digit 1. If the assumption of uniform distribution holds, this justifies equation (2.6) that $\text{Prob}(D_1 = 1) = \log_{10}(2) - \log_{10}(1) \approx 0.301$ When $\log_B(s(n))$ is uniformly distributed mod 1 for all B, then $s(n)$ conforms to Benford's Law.

In order to prove Theorem 2, we just need to show that if (x_n) is good, $\log(x_n)$ is uniformally distributed mod 1. In this case, we still need the following two theorems to support our proof.

Theorem 3 (Weyl's Criterion). *The sequence $\omega = (x_n), n = 1, 2, \ldots,$ is u.d. mod 1 if and only if*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i h x_n} = 0 \textit{ for all integers } h \neq 0. \tag{2.8}$$

Proof. The necessity follows from Corollary 1.1. Suppose that (x_n) fulfills the criterion (2.8). We can show that (2.4) is true for every complex-valued continuous function f on $\mathbb{R}$ with period 1. Let $\epsilon > 0$ be an arbitrary number. According to the Weierstrass approximation theorem, there exists a trigonometric polynomial $\Psi(x)$, that is, a finite linear combination of functions with the term $e^{2\pi i h x_n}$, $h \in \mathbb{Z}$, with complex coefficients, such that

$$\sup_{0\le x\le 1} \big| f(x) - \Psi(x) \big| \le \epsilon. \tag{2.9}$$

$$\left|\int_0^1 f(x)dx - \frac{1}{N}\sum_{n=1}^{N} f(x_n)\right| \le \left|\int_0^1 (f(x)-\Psi(x))dx\right| + \left|\int_0^1 \Psi(x)dx - \frac{1}{N}\sum_{n=1}^{N}\Psi(x_n)\right|$$

$$+\left|\frac{1}{N}\sum_{n=1}^{N}(f(x_n)-\Psi(x_n))\right|.$$

According to (2.9), the first term and third term on the right both are $\le \epsilon$ for any value of N. As we assume (2.8) holds at first, if N is sufficiently large, the second term on the right is $\le \epsilon$. ■

Theorem 4 (Theorem 3.5 in [7]). *Let $f(x)$ be a function defined for $x \ge 1$ which is k-times differentiable for all $x \ge x_0$ for some $x_0 \in \mathbb{R}_+$, $k \in \mathbb{N}$. Suppose that $f^{(k)}$ is eventually monotonic,*

$$\lim_{x\to\infty} f^{(k)}(x) = 0, \tag{2.10}$$

and

$$\lim_{x\to\infty} x\left|f^{(k)}(x)\right| = \infty, \tag{2.11}$$

then the sequence $\{f(n) : n \in \mathbb{N}\}$ is uniformly distributed mod 1.

In order to establish proof of Theorem 4, we need the following theorems and corollary as support.

Theorem 5 (Theorem 2.5 in [7]). *A sequence of real numbers $f(n), n = 1, 2, \dots,$ satisfies that $\Delta f(n) = f(n+1) - f(n)$ is monotone as n increases. Therefore, if*

$$\lim_{n\to\infty} \Delta f(n) = 0 \textit{ and } \lim_{n\to\infty} n\left|\Delta f(n)\right| = \infty. \tag{2.12}$$

Then the sequence $f(n)$ is u.d. mod 1.

Proof. For every pair of real numbers u and v,

$$\left|\frac{e^{2\pi iu}-e^{2\pi iv}-2\pi i(u-v)e^{2\pi iv}}{e^{2\pi iv}}\right| = \left|e^{2\pi i(u-v)}-1-2\pi i(u-v)\right|$$

$$= 4\pi^2\left|\int_0^{u-v}(u-v-w)e^{2\pi iw}dw\right|$$

$$\leq 4\pi^2\left|\int_0^{u-v}(u-v-w)dw\right|$$

$$= 2\pi^2(u-v)^2. \tag{2.13}$$

If we let $u = hf(n+1)$ and $v = hf(n), h \in \mathbb{N}$. Then, it follows from (2.13) that:

$$\left|\frac{e^{2\pi ihf(n+1)}}{\Delta f(n)}-\frac{e^{2\pi ihf(n)}}{\Delta f(n)}-2\pi ihe^{2\pi ihf(n)}\right| \leq 2\pi^2h^2\left|\Delta f(n)\right| \text{ for } n \geq 1$$

$$\left|\frac{e^{2\pi ihf(n+1)}}{\Delta f(n+1)}-\frac{e^{2\pi ihf(n)}}{\Delta f(n)}-2\pi ihe^{2\pi ihf(n)}\right|$$

$$\leq\left|\frac{1}{\Delta f(n)}-\frac{1}{\Delta f(n+1)}\right|+2\pi^2h^2\left|\Delta f(n)\right| \text{ for } n \geq 1 \tag{2.14}$$

$$\left|2\pi ih\sum_{n=1}^{N-1}e^{2\pi ihf(n)}\right| = \left|\sum_{n=1}^{N-1}(2\pi ihe^{2\pi ihf(n)}-\frac{e^{2\pi ihf(n+1)}}{\Delta f(n+1)}+\frac{e^{2\pi ihf(n)}}{\Delta f(n)})+\frac{e^{2\pi ihf(N)}}{\Delta f(N)}-\frac{e^{2\pi ihf(1)}}{\Delta f(1)}\right|$$

$$\leq\sum_{n=1}^{N-1}\left|2\pi ihe^{2\pi ihf(n)}-\frac{e^{2\pi ihf(n+1)}}{\Delta f(n+1)}+\frac{e^{2\pi ihf(n)}}{\Delta f(n)}\right|+\frac{1}{\left|\Delta f(N)\right|}+\frac{1}{\left|\Delta f(1)\right|}$$

$$\leq\sum_{n=1}^{N-1}\left|\frac{1}{\Delta f(n)}-\frac{1}{\Delta f(n+1)}\right|+2\pi^2h^2\sum_{n=1}^{N-1}\left|\Delta f(n)\right|+\frac{1}{\left|\Delta f(N)\right|}+\frac{1}{\left|\Delta f(1)\right|}.$$

Because $\Delta f(n)$ is monotone, we then have

$$\left|\frac{1}{N}\sum_{n=1}^{N-1}e^{2\pi ihf(n)}\right| \leq \frac{1}{\pi|h|}(\frac{1}{N\left|\Delta f(1)\right|}+\frac{1}{N\left|\Delta f(N)\right|})+\frac{\pi|h|}{N}\sum_{n=1}^{N-1}\left|\Delta f(n)\right|.$$

If (2.12) is true, then

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N-1}e^{2\pi i h f(n)}=0.$$

We recall Weyl's Criterion that the sequence $f(n)$ which satisfies the above equation is u.d. mod 1.

■

Corollary 5.1 (Fejer's Theorem). *Let $f(x)$ be a function defined for $x \geq 1$ which is differentiable for $x \geq x_0$. If when $x \to \infty$, $f'(x)$ monotonically tends to 0, and $n\big|f'(x)\big|$ goes to ∞, then the sequence $f(n), n = 1, 2, \dots,$ is u.d. mod 1.*

Proof. The mean value theorem yields that there is a point c in the interval $(n, n+1)$ of the function $f(x)$ that

$$f'(c)=\frac{f(n+1)-f(n)}{(n+1)-n}=\Delta f(n).$$

So when $x \to \infty$ and the conditions for Fejer's Theorem holds, then we get $\Delta f(n) \to 0, x\big|\Delta f(n)\big| \to \infty$, which satisfies the conditions of Theorem 5.

■

Theorem 6. *For each positive integer h, a sequence $(f(n+h)-f(n))$ with real numbers is u.d. mod 1, if $\Delta^k f(n)$ is monotone in n and as $n \to \infty$, we have $\Delta^k f(n) \to 0$ and $n\big|\Delta^k f(n)\big| \to \infty$ for $k \in \mathbb{N}$.*

Proof. When $k = 1$, we have Theorem 5. For $k > 1$, we first assume the theorem is true. Let a sequence $(f(n))$ with $\lim_{n\to\infty}\Delta^{(k+1)}f(n) = 0$, $\lim_{n\to\infty} n\Big|\Delta^{(k+1)}f(n)\Big| = \infty$, and $\Delta^{(k+1)}f(n)$ is monotone in n. For a fixed positive integer h,

$$f(n+h)-f(n)=\Sigma_{j=0}^{h-1}\Delta f(n+j)$$

$$\Delta^{(k)}(f(n+h)-f(n))=\Sigma_{j=0}^{h-1}\Delta^{(k+1)}f(n+j)$$

Then we have $\Delta^{(k+1)}(f(n+h)-f(n))$ is monotone in n, $\lim_{n\to\infty}\Delta^{(k+1)}(f(n+h)-f(n))=0$, and $\lim_{n\to\infty} n\Big|\Delta^{(k+1)}(f(n+h)-f(n))\Big|=\infty$. By induction hypothesis, we prove that the sequence $(f(n+h)-f(n))$ is u.d. mod 1.

■

With Theorem 6, we can extend Fejer's Theorem as following. Since $\Delta^k f(n)=f^{(k)}(c)$ and $n\big|\Delta^k f(n)\big|=xf^{(k)}(c)$, $\lim_{n\to\infty}\Delta^{(k)}f(n)=0$ and $\lim_{n\to\infty} n\Big|\Delta^{(k)}f(n)\Big|=\infty$ are equivalent to $\lim_{n\to\infty} f^{(k)}(x)=0$ and $\lim_{n\to\infty} x\Big|f^{(k)}(x)\Big|=\infty$. Thus, if $f^{(k)}(x)$ tends to 0 and $x\Big|f^{(k)}(x)\Big|$ tends to ∞ monotonically as $n\to\infty$, $(f(n+h)-f(n))$ is u.d. mod 1 for any $h\geq 1$. In order to prove that $(f(n))$ is u.d. mod 1, we are going to introduce Van der Corput's Fundamental Inequality and Van der Corput's Difference Theorem.

Lemma 7.1 (Van der Corput's Fundamental Inequality)**.** *Let $u_1,\dots,u_N\in\mathbb{C}$, and let H be an integer that $1\leq H\leq N$. Then*

$$H^2\left|\sum_{n=1}^{N}u_n\right|^2\leq H(N+H-1)\sum_{n=1}^{N}|u_n|^2+2(N+H-1)\sum_{h=1}^{H-1}(H-h)Re\sum_{n=1}^{N-h}u_n\bar{u}_{n+h},\qquad(2.15)$$

where Re z is the real part of the a complex number z.

Proof. Let $u_n=0$ for $n\leq 0$ and $n>N$.

$$H\sum_{n=1}^{N}u_n=\sum_{p=1}^{N+H-1}\sum_{h=0}^{H-1}u_{p-h}\qquad(2.16)$$

$$H^2\left|\sum_{n=1}^{N}u_n\right|^2\leq(N+H-1)\sum_{p=1}^{N+H-1}\left|\sum_{h=0}^{H-1}u_{p-h}\right|^2$$

$$=(N+H-1)\sum_{p=1}^{N+H-1}(\sum_{r=0}^{H-1}u_{p-r})(\sum_{s=0}^{H-1}\bar{u}_{p-s})$$

$$= (N+H-1)\sum_{p=1}^{N+H-1}\sum_{h=0}^{H-1}\left|u_{p-h}\right|^2 + 2(N+H-1)Re\sum_{p=1}^{N+H-1}\sum_{r,s=0,s<r}^{H-1}u_{p-r}\bar{u}_{p-s}$$

$$= (N+H-1)(\Sigma_1 + 2Re\Sigma_2),$$

where $\Sigma_1 = H\sum_{n=1}^{N}\left|u_n\right|^2, \Sigma_2 = \sum_{n,h}u_n\bar{u}_{n+h}, n = 1,2,\ldots N$ and $h = r-s = 1,2,\ldots,H-1 (s < r)$

For a fixed n that $1 \leq n \leq N$ and fixed h that $1 \leq h \leq H-1$, the possible combinations for (r,s) will be $(h,0),(h+1,1),\ldots,(H-1,H-h-1)$. Each of them gives a unique p. We will have $H-h$ repetitions of $u_n\bar{u}_{n+h}$ in Σ_2.

$$\Sigma_2 = \sum_{h=1}^{H-1}(H-h)\sum_{n=1}^{N}u_n\bar{u}_{n+h}$$

Because for $n > N, u_n = 0$, the summation over n could be altered to $1 \leq n \leq N-h$.

■

Theorem 7 (Van der Corput's Difference Theorem). *Given a sequence of real numbers (x_n). If for every positive integer h, the sequence $(x_{n+h}-x_n), n = 1,2,\ldots,$ is u.d. mod 1, then (x_n) is also u.d. mod 1.*

Proof. Let $u_n = e^{2\pi i m x_n}, m$ is a constant nonzero integer. We divide (2.15) in Lemma 7.1 by H^2N^2.

$$\left|\frac{1}{N}\sum_{n=1}^{N}e^{2\pi i m x_n}\right|^2 \leq \frac{N+H-1}{HN} + 2\sum_{h=1}^{H-1}\frac{(N+H-1)(H-h)(N-h)}{H^2N^2}\left|\frac{1}{N-h}\sum_{n=1}^{N-h}e^{2\pi i m(x_n - x_{n+h})}\right|.$$

For each $h \geq 1, x_n - x_{n+h}$ is u.d. mod 1, now we have,

$$\lim_{N\to\infty}\frac{1}{N-h}\sum_{n=1}^{N-h}e^{2\pi i m(x_n-x_{n+h})} = 0, \text{ for each } h \geq 1.$$

Combining the above two equations:

$$\overline{\lim_{N\to\infty}}\left|\frac{1}{N}\sum_{n=1}^{N}e^{2\pi imx_n}\right|^2 \leq \frac{1}{H}.$$

It holds for every H, if we take H as large as possible, then we have

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}e^{2\pi imx_n}=0.$$

■

With Theorem 5, Corollary 5.1, Theorem 6, Lemma 6.1, we are able to prove Theorem 4 by induction.

Proof. For $k=1$, it is the same to prove Corollary 5.1, which we already did.
Then, for $k>1$, the extension of Fejer's Theorem indicates that if the conditions for Theorem 4 hold, then the sequence $f(n+h)-f(n)$ is u.d. mod 1. By Van der Corput's Difference Theorem, the sequence is proved to be $f(n)$ is u.d. mod 1 [7]. ■

Through the definition of uniform distribution mod 1 and Weyl's Criterion, the following Lemmas could be derived from Theorem 4:

Lemma 4.1. *If $f(n)$ is Benford and $f(n)\sim g(n)$, then $g(n)$ is Benford.*

Now, we are ready to prove Theorem 2.

Proof. Let an integer-valued function $x(n)$ be good first. It follows that $x(n)\sim a(n)e^{b(n)}$. According to Lemma 4.1, in order to show $x(n)$ is Benford, it is sufficient to show $a(n)e^{b(n)}$ is Benford. This is also to prove $\log a(n)e^{b(n)} = \log(a(n))+b(n)(mod 1)$ is uniformly distributed. By definition of a good function, we have equation (2.9). It satisfies the condition (2.14) that $b(n)$ is uniformly distributed mod 1. From (2.10), the limit will not be affected if we add $\log(a(n))$ to $b(n)$ [1]. ■

3 Established Cases of Benford's Law

In the first section, we mentioned that Frank Benford experimented on 20 different data sets, which included some mathematical sequences such as $n^{-1}, \sqrt{n}, n!$, etc. His work inspired a lot of other researchers to further investigate sequences. Many sequences have been proven Benford. In this section, we include two established cases, their proofs, and how Theorem2 has been applied in each case.

3.1 $n!$

n	**n!**
0	1
1	1
2	2
3	6
4	24
5	120
6	720
7	5040
8	40320
9	36880
10	3628800

Table 7: First 20 Terms of the Factorial Sequence $n!$

The sequence $n!$ is a Benford sequence. The proof is rather simple, followed by Theorem 2.

Proof. Recall Stirling's formula:

$$n! \sim \frac{1}{\sqrt{2\pi}} n^{n+\frac{1}{2}} e^{-n} \tag{3.1}$$

By Definition 3, the sequence $n!$ is good. Hence, $n!$ is Benford from Theorem 2 [6]. ■

3.2 Partition Functions

The partition function $p(n)$ calculates how many different ways that the integer n could be written as a sum of positive integers. We consider that the order of the addends does not matter. For example, the integer 5 could be written as follows:

$$
\begin{gathered}
5=5\\
5=4+1\\
5=3+2\\
5=3+1+1\\
5=2+2+1\\
5=2+1+1+1\\
5=1+1+1+1+1.
\end{gathered}
$$

Thus, $p(5) = 7$ and such function $p(n)$ is also called an unrestricted partition function [10]. The following Table 8 and 9 gives $p(n)$ for $n = 1, 2, \ldots, 50$ and the digit frequency.

n	**p(n)**	**n**	**p(n)**	**n**	**p(n)**
0	1	**19**	490	**38**	26015
1	1	**20**	627	**39**	31185
2	2	**21**	792	**40**	37338
3	3	**22**	1002	**41**	44583
4	5	**23**	1255	**42**	53174
5	7	**24**	1575	**43**	63261
6	11	**25**	1958	**44**	75175
7	15	**26**	2436	**45**	89134
8	22	**27**	3010	**46**	105558
9	30	**28**	3718	**47**	124754
10	42	**29**	4565	**48**	147273
11	56	**30**	5604	**49**	173525
12	77	**31**	6842		
13	101	**32**	8349		
14	135	**33**	10143		
15	176	**34**	12310		
16	231	**35**	14883		
17	297	**36**	17977		
18	385	**37**	21637		

Table 8: First 50 Terms of the Partition Function $p(n)$

Digit	Frequency
1	19
2	7
3	7
4	4
5	4
6	3
7	4
8	2
9	0

Table 9: Digit Frequency of $p(n), n = 1, 2, \ldots, 50$

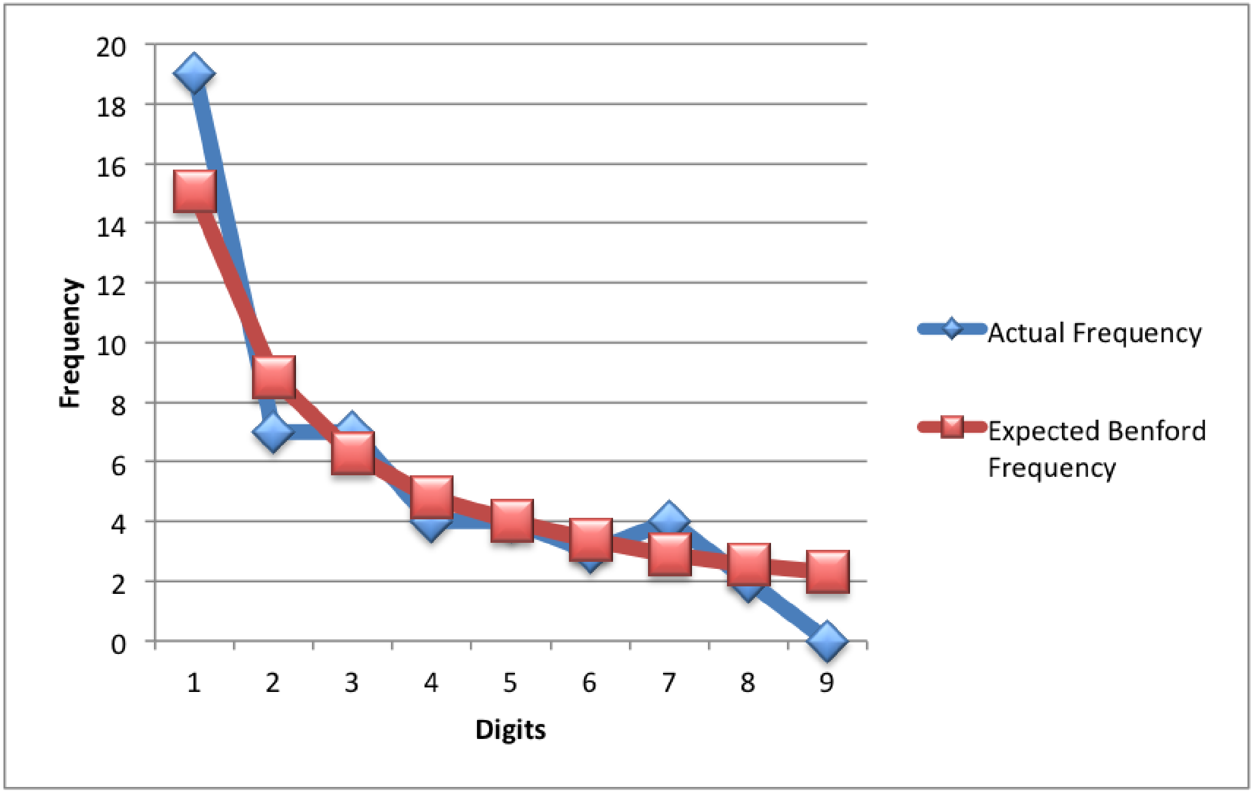

Figure 4: First Digit Frequency of $p(n), n = 1, 2, \ldots, 50$

We could observe from Figure 4 that except digit 1 and 9, the first digit distribution accord with the expected one. To prove the partition function is actually a Benford sequence mathematically, we recall that Theorem 2 states that: if a sequence (x_n) is good, then it is Benford.

Corollary 2.1. *The partition function $p(n)$ is Benford.*

Proof. Hardy and Ramanujn obtain the following asymptotic with the circle method and modular functions[10]:

$$p(n) \sim \frac{1}{4n\sqrt{3}} e^{\pi\sqrt{2n/3}}. \tag{3.2}$$

It is obvious that $p(n)$ is good and hence Benford [1]. ■

Bibliography

[1] Theresa Anderson, Larry Rolen, and Ruth Stoehr. Benford's law for coefficients of modular forms and partition functions. *Proceedings of the American Mathematical Society*, 139(5):1533–1541, 2011.

[2] Frank Benford. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, pages 551–572, 1938.

[3] Charles APN Carslaw. Anomalies in income numbers: Evidence of goal oriented behavior. *Accounting Review*, pages 321–327, 1988.

[4] United Nations Statistics Division. Gdp and its breakdown at current prices in us dollars (all countries for all years - sorted alphabetically), 2014.

[5] Cindy Durtschi, William Hillison, and Carl Pacini. The effective use of benford's law to assist in detecting fraud in accounting data. *Journal of forensic accounting*, 5(1):17–34, 2004.

[6] Adrien Jamain. Benford's law. *Unpublished Dissertation Report, Department of Mathematics, Imperial College, London*, 2001.

[7] Lauwerens Kuipers and Harald Niederreiter. *Uniform distribution of sequences.* Courier Corporation, 2012.

[8] Simon Newcomb. Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, 4(1):39–40, 1881.

[9] Mark Nigrini. *Benford's Law: Applications for forensic accounting, auditing, and fraud detection*, volume 586. John Wiley & Sons, 2012.

[10] Eric W Weisstein. Partition function p. 2002.