Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

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
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

1 Introduction ................................................................. 1
   1.1 Historical Background ............................................. 4
   1.2 Real-life Applications ............................................. 7
2 Mathematical Foundations of Benford’s Law .................... 11
   2.1 Uniform Distribution Modulo 1 .................................. 11
   2.2 Definition of “Benford’s Law Base B” ......................... 14
   2.3 Mathematical Justification of Benford’s Law ................. 16
3 Established Cases of Benford’s Law ............................... 25
   3.1 \( n! \) ................................................................. 25
   3.2 Partition Functions ................................................ 26
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>First Digit Graph of the Data in Table 1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Expected First and Second Digits Distribution Derived by Newcomb</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>Logarithm Scale Bar</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>First Digit Frequency of $p(n), n = 1, 2, \ldots, 50$</td>
<td>28</td>
</tr>
</tbody>
</table>
## List of Tables

2. Digit Frequency Table of the Data in Table 1 . . . . . . . . . . . . . . . . . . 4
3. Expected Digits Frequency Derived By Newcomb . . . . . . . . . . . . . . . 5
4. Benford’s Analysis of Natural Data Sets in 1938 [2] . . . . . . . . . . . . . . 7
5. First 12 Distinct Numbers in Fibonacci sequence with Different Bases 10,2,5 15
6. Logarithms of Numbers 1-10 in Base 10 . . . . . . . . . . . . . . . . . . . . . 17
7. First 20 Terms of the Factorial Sequence \( n! \) . . . . . . . . . . . . . . . . . 25
8. First 50 Terms of the Partition Function \( p(n) \) . . . . . . . . . . . . . . . . 27
9. Digit Frequency of \( p(n), n = 1, 2, \ldots, 50 \) . . . . . . . . . . . . . . . 27
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
1. **INTRODUCTION**

completely different story.

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

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 1\textsuperscript{st} ranked country (the US) and the 60\textsuperscript{th} 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.

<table>
<thead>
<tr>
<th>Digit</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>15</td>
<td>21</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Proportion</td>
<td>0.25</td>
<td>0.35</td>
<td>0.1</td>
<td>0.1</td>
<td>0.083</td>
<td>0.033</td>
<td>0.0167</td>
<td>0.05</td>
<td>0.0167</td>
</tr>
</tbody>
</table>

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

![First Digit Graph](image)

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 19th 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: 
\[
\text{Prob}(D_1 = N) = \log_{10}(N + 1) - \log_{10}(N).
\]
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]:

<table>
<thead>
<tr>
<th>( N )</th>
<th>( \text{Prob}(D_1 = N) )</th>
<th>( \text{Prob}(D_2 = N) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-</td>
<td>0.1197</td>
</tr>
<tr>
<td>1</td>
<td>0.3010</td>
<td>0.1139</td>
</tr>
<tr>
<td>2</td>
<td>0.1761</td>
<td>0.1088</td>
</tr>
<tr>
<td>3</td>
<td>0.1249</td>
<td>0.1043</td>
</tr>
<tr>
<td>4</td>
<td>0.0969</td>
<td>0.1003</td>
</tr>
<tr>
<td>5</td>
<td>0.0792</td>
<td>0.0967</td>
</tr>
<tr>
<td>6</td>
<td>0.0669</td>
<td>0.0934</td>
</tr>
<tr>
<td>7</td>
<td>0.0580</td>
<td>0.0904</td>
</tr>
<tr>
<td>8</td>
<td>0.0512</td>
<td>0.0876</td>
</tr>
<tr>
<td>9</td>
<td>0.0458</td>
<td>0.0850</td>
</tr>
</tbody>
</table>

Table 3: Expected Digits Frequency Derived By Newcomb

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

Completely deviating from the intuition: \( \text{Prob}(D_i = N) = \frac{1}{9}, \text{for} \ N = 0, 1, \ldots, 9, i \in \mathbb{N}, \) except \( \text{Prob}(D_1 = 0) \) 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.
### 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 Charlier 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. 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 Benford 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.

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.

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 $x \pmod{1}$, $\{x\} \in$ 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 [1]). 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

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

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} c_{[a,b)}(\{x_n\}) = \int_{0}^{1} c_{[a,b)}(x)dx \quad (2.2)
\]

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

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} f(\{x_n\}) = \int_{0}^{1} f(x)dx \quad (2.3)
\]

**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:

\[
\int_{0}^{1} (f(x) - f_1(x))dx \leq \epsilon
\]
\[
\int_0^1 f(x)dx - \epsilon \leq \int_0^1 f_1(x)dx = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} f_1(\{x_n\}) \\
\leq \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} f(\{x_n\}) \leq \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} f(\{x_n\}) \\
\leq \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} f_2(\{x_n\}) = \int_0^1 f_2(x)dx \leq \int_0^1 f(x)dx + \epsilon.
\]

Since \(\epsilon\) 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 \(I\), 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:

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

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

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

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

Since \(\epsilon\) 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*

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} f(x_n) = \int_0^1 f(x)dx.
\]  

(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 \(f(\{x_n\}) = \)
2. MATHEMATICAL FOUNDATIONS OF BENFORD’S LAW

\( f(x_n), \) 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 

\[ \blacksquare \]

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, \( B \in \mathbb{N}^* = \{1, 2, \ldots \} \).

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, \ldots, 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 \( j, \) \( x_k = 1, 2, \ldots, B - 1. \) 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.
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

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

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

\[
\lim_{N \to \infty} B(d, N, B; \omega) \equiv \log_B(d + 1) - \log_B(d) \pmod{1}.
\]

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 \( 100_2 = 4_{10} \), \((\log_2(5) - \log_2(4)) \pmod{1} \approx 0.3219\), 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.

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

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

\[ \text{Prob}(D_1D_2 = N_1N_2) = \log_B(N_1N_2 + 1) - \log_B(N_1N_2). \]

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

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

(2)

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

(3)

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

<table>
<thead>
<tr>
<th>Digits N</th>
<th>( \log_{10}(N) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.301</td>
</tr>
<tr>
<td>3</td>
<td>0.477</td>
</tr>
<tr>
<td>4</td>
<td>0.602</td>
</tr>
<tr>
<td>5</td>
<td>0.698</td>
</tr>
<tr>
<td>6</td>
<td>0.778</td>
</tr>
<tr>
<td>7</td>
<td>0.845</td>
</tr>
<tr>
<td>8</td>
<td>0.903</td>
</tr>
<tr>
<td>9</td>
<td>0.954</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>

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 uniformly 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

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

**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

$$
\sup_{0 \leq x \leq 1} |f(x) - \Psi(x)| \leq \epsilon. \tag{2.9}
$$
\[ \left| \int_0^1 f(x)dx - \frac{1}{N} \sum_{n=1}^{N} f(x_n) \right| \leq \left| \int_0^1 (f(x) - \Psi(x))dx \right| + \left| \int_0^1 \Psi(x)dx - \frac{1}{N} \sum_{n=1}^{N} \Psi(x_n) \right| \]
\[ + \left| \frac{1}{N} \sum_{n=1}^{N} (f(x_n) - \Psi(x_n)) \right| . \]

According to (2.9), the first term and third term on the right both are \( \leq \epsilon \) 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 \( \leq \epsilon \). \[ \blacksquare \]

**Theorem 4** (Theorem 3.5 in [7]). Let \( f(x) \) be a function defined for \( x \geq 1 \) which is \( k \)-times differentiable for all \( x \geq x_0 \) 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 \( f(n), n = 1, 2, \ldots, \) satisfies that \( \Delta f(n) = f(n + 1) - f(n) \) is monotone as \( n \) increases. Therefore, if

\[ \lim_{n \to \infty} \Delta f(n) = 0 \ and \ \lim_{n \to \infty} n|\Delta f(n)| = \infty. \tag{2.12} \]

Then the sequence \( f(n) \) is u.d. mod 1.
Proof. For every pair of real numbers \( u \) and \( v \),

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

\[
= 4\pi^2 \left| \int_0^{u-v} (u - v - w)e^{2\pi iu} \, dw \right|
\]

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

\[
= 2\pi^2 (u - v)^2. \quad (2.13)
\]

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

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

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

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

\[
2\pi i h \sum_{n=1}^{N-1} e^{2\pi ihf(n)} = \sum_{n=1}^{N-1} \left( 2\pi ihe^{2\pi ihf(n)} \right) - \frac{e^{2\pi ihf(n+1)}}{\Delta f(n+1)} + \frac{e^{2\pi ihf(n)}}{\Delta f(n)} + \frac{e^{2\pi ihf(N)}}{\Delta f(N)} - \frac{e^{2\pi ihf(1)}}{\Delta f(1)}
\]

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

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

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

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

\[\Box\]

**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 \( n|f'(x)| \) goes to \( \infty \), then the sequence \( f(n), n = 1, 2, \ldots \), 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 \( \Delta f(n) \to 0, x|\Delta f(n)| \to \infty \), which satisfies the conditions of Theorem 5.

\[\Box\]

**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 \( n|\Delta^k f(n)| \to \infty \) 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 \( \lim_{n \to \infty} \Delta^{(k+1)} f(n) = 0 \), \( \lim_{n \to \infty} n|\Delta^{(k+1)} f(n)| = \infty \), and \( \Delta^{(k+1)} f(n) \) is monotone in \( n \). For a fixed positive integer \( h \),
\[
f(n + h) - f(n) = \sum_{j=0}^{h-1} \Delta f(n + j)
\]
\[
\Delta^{(k)}(f(n + h) - f(n)) = \sum_{j=0}^{h-1} \Delta^{(k+1)} f(n + j)
\]
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 \( \lim_{n \to \infty} n \left| \Delta^{(k+1)}(f(n+h) - f(n)) \right| = \infty \). 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 \( n|\Delta^k f(n)| = xf^{(k)}(c) \), \( \lim_{n \to \infty} \Delta^k f(n) = 0 \) and \( \lim_{n \to \infty} n|\Delta^k f(n)| = \infty \) are equivalent to \( \lim_{n \to \infty} f^{(k)}(x) = 0 \) and \( \lim_{n \to \infty} x|f^{(k)}(x)| = \infty \). Thus, if \( f^{(k)}(x) \) tends to 0 and \( x|f^{(k)}(x)| \) tends to \( \infty \) 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 \( u_1, \ldots, u_N \in \mathbb{C} \), and let \( H \) be an integer that \( 1 \leq H \leq N \). Then

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

(2.15)

where \( \text{Re} z \) is the real part of the a complex number \( z \).

**Proof.** Let \( u_n = 0 \) for \( n \leq 0 \) and \( n > N \).

\[
H \sum_{n=1}^{N} u_n = \sum_{p=1}^{N-H+1} \sum_{h=0}^{H-1} u_{p-h}.
\]  

(2.16)

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

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

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

where \( \Sigma_1 = H \sum_{n=1}^{N} |u_n|^2 \), \( \Sigma_2 = \sum_{n,h} u_n \bar{u}_{n+h}, n = 1, 2, \ldots N \) 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 \( \Sigma_2 \).

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

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

\[\boxed{}\]

**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 \( H^2 N^2 \).

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

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:

\[
\lim_{N \to \infty} \left| \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i m x_n} \right|^2 \leq \frac{1}{H}.
\]

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

\[
\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} e^{2\pi i m x_n} = 0.
\]

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 Fejér’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 \( \log a(n)e^{b(n)} = \log(a(n)) + b(n)(mod1) \) 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 Theorem 2 has been applied in each case.

3.1 $n!$

<table>
<thead>
<tr>
<th>$n$</th>
<th>$n!$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
</tr>
<tr>
<td>5</td>
<td>120</td>
</tr>
<tr>
<td>6</td>
<td>720</td>
</tr>
<tr>
<td>7</td>
<td>5040</td>
</tr>
<tr>
<td>8</td>
<td>40320</td>
</tr>
<tr>
<td>9</td>
<td>36880</td>
</tr>
<tr>
<td>10</td>
<td>3628800</td>
</tr>
</tbody>
</table>

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:

$$n! \sim \frac{1}{\sqrt{2\pi}} n^{n + \frac{1}{2}} e^{-n}$$  \hspace{1cm} (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:

\[
5 = 5 \\
5 = 4 + 1 \\
5 = 3 + 2 \\
5 = 3 + 1 + 1 \\
5 = 2 + 2 + 1 \\
5 = 2 + 1 + 1 + 1 \\
5 = 1 + 1 + 1 + 1 + 1.
\]

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.
<table>
<thead>
<tr>
<th>Digit</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 9: Digit Frequency of $p(n)$, $n = 1, 2, \ldots, 50$
3. ESTABLISHED CASES OF BENFORD’S LAW

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]\:

\[
p(n) \sim \frac{1}{4n\sqrt{3}} e^{\pi\sqrt{2n/3}}.
\]

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


