UNIVERSITY OF DUNDEE

MASTER THESIS

BENFORD'S LAW

Author:
Sana NAZ

Supervisor:
Dr. Agis ATHANASSOULIS

*A thesis submitted in fulfillment of the requirements
for the degree of MS.c Applied Mathematics*

August 30, 2024

"In our lust for measurment, we can frequentlt measure that which we can rather than that which we wish to measure... and forget that there is a difference."

George Udny Yule

UNIVERSITY OF DUNDEE

Abstract

School of Science and Engineering
Department of Mathematics

MS.c Applied Mathematics

BENFORD'S LAW

by Sana NAZ

This project explores Benford's Law, a statistical principle describing the frequency distribution of leading digits in various datasets. The study begins by explaining Benford's Law and providing examples of datasets that comply, such as population figures and financial records, as well as those that do not, such as sequential or uniformly distributed numbers. Statistical techniques, including chi-square hypothesis testing, are employed to measure the adherence of datasets to Benford's Law, demonstrated through real-life examples like census data. Additionally, the project examines the scale invariance of Benford's Law and explores transformations that preserve this behavior. Finally, the practical applications of Benford's Law in fields such as finance and statistics are discussed, highlighting its relevance and utility. The aim is to provide a comprehensive understanding of Benford's Law, its conditions, and its broad applicability.

Acknowledgements

I would like to express my sincere gratitude to all those who supported me throughout this project. Firstly, I am deeply grateful to my supervisor, Dr. Agis Athanassoulis, for their invaluable guidance, encouragement, and insightful feedback, which were instrumental in the successful completion of this project.

Special thanks to the [Department of Mathematics] at University of Dundee for providing the resources and environment conducive to conducting this research.

I am also thankful to my family and friends for their unwavering support and understanding during the entire duration of this project. Your encouragement kept me motivated throughout.

Finally, I acknowledge the authors and researchers whose work laid the foundation for this study, and I am grateful for the wealth of knowledge and insights shared in the field of Benford's Law.

Thank you all for your contributions and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In statistics Benford's Law describes the frequency distribution of leading digits in many naturally occurring datasets, revealing that smaller digits occur more often as the first digit than larger ones. However, it is important to note that Benford's Law does not apply to all datasets, and its applicability depends on the nature and range of the data.Before describing the law, we need to establish some notation.

In secondary school, you likely learned about scientific notation: any positive number x can be expressed as $S(x)10^k$, where $S(x)[1, 10)$ is the significand, and k is an integer known as the exponent. The integer part of the significand is called the leading digit, or the first digit.While some might refer to S(x) as the mantissa, this can be misleading.

1.1 History of Benford's Law

Though it is known as Benford's Law, the first observation of this digit bias was actually made by the astronomer-mathematician Simon Newcomb, more than 50 years before Benford. Newcomb, born in Nova Scotia in 1835 and passing away in Washington, DC in 1909, made his observation public in 1881 through a brief article in the American Journal of Mathematics titled "Note on the Frequency of Use of the Different Digits in Natural Numbers."

Newcomb's article begins with the observation:

"That the ten digits do not occur with equal frequency must be evident to anyone making much use of logarithmic tables and noticing how much faster the first pages wear out than the last ones. The first significant figure is oftener 1 than any other digit, and the frequency diminishes up to 9. The question naturally arises whether the reverse would be true of logarithms. That is, in a table of anti-logarithms, would the last part be more used than the first, or would every part be used equally? The law of frequency in the one case may be deduced from that in the other. The question we have to consider is, what is the probability that if a natural number be taken at random its first significant digit will be n, its second n' etc." [9]

Newcomb suggests that natural numbers should be viewed as ratios of quantities rather than isolated values. Therefore, instead of selecting a number at random, he proposes selecting two numbers and then determining the probability that the first significant digit of their ratio is a given digit n. By forming an indefinite number of such ratios independently and examining their quotients, one can find the limit toward which the probability of the first digit being n approaches.

In this concise article, Newcomb identifies two significant properties of digit distribution:

- Unequal Likelihood of Digits: Not all digits are equally likely y to appear as the first digit of natural numbers.

- Importance of Scale: The numerical value of a physical quantity is dependent on the scale used, leading Newcomb to suggest that the correct subjects of study are ratios of measurements.

Newcomb concludes with a quantification of this bias, noting:
"The law of probability of the occurrence of numbers is such that all mantissa of their logarithms are equally probable."
He provides a table (i.e,Table 1.1) that details the probabilities of observing each digit as the first or second digit.

d	Probability first digit d	Probability second digit d
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 1.1: Newcomb's probabilities of first or second digit

This table illustrates Newcomb's conjecture regarding the probabilities of observing specific digits as the first or second digit in a set of natural numbers.

Benford's Law which is also called the First Digit Law describe In 1938, Benford published "The Law of Anomalous Numbers," which became known as Benford's Law. Frank Benford, a physicist at General Electric, Born on July 10, 1883, in Johnstown, Pennsylvania, Benford was a notable inventor and researcher in light and optics. Despite his professional focus, he had a keen interest in mathematics.

He also presents some justification:
" It has been noted that the pages of a frequently used table of common logarithms show signs of selective usage. The pages containing the logarithms for the lower numbers, such as 1 and 2, tend to be more worn and stained compared to those for the higher numbers, like 8 and 9. While the condition of a logarithm table might not seem particularly significant, this phenomenon becomes more intriguing when we consider that these tables play a crucial role in the development of our scientific, engineering, and factual literature. The relative wear and tear on these pages may offer insights into how we think and respond when dealing with numerical information."[1]

To test his hypothesis, Benford analyzed 20,229 records from 20 diverse datasets, including newspapers and scientific constants. He found that the number 1 appeared as the first digit about 30.1% of the time, while the number 9 appeared only 4.6% of the time, following a logarithmic distribution. If the digits 1 through 9 had an equal probability, each would occur 11.1% of the time. However, in many real-world datasets, this is not the case.

Definition 1.1.1 *A set of numbers is said to satisfy Benford's law if the leading digit d, $(d\epsilon 1, ..., 9)$ occurs with probability.*

$$prob(d) = log_{10}(d+1) - log_{10}(d) = log_{10}(\frac{d+1}{d}) = log_{10}(1+\frac{1}{d})$$

Benford's Law predicts that the probability of a randomly chosen number in a dataset starting with a specific digit d is given by the formula $log_{10}(1+\frac{1}{d})$ If the actual distribution of leading digits in a dataset closely matches these predicted probabilities, the dataset is often described as "Benford." For example, the areas of countries or the populations of counties might be referred to as "Benford" if their leading digits conform to this distribution.

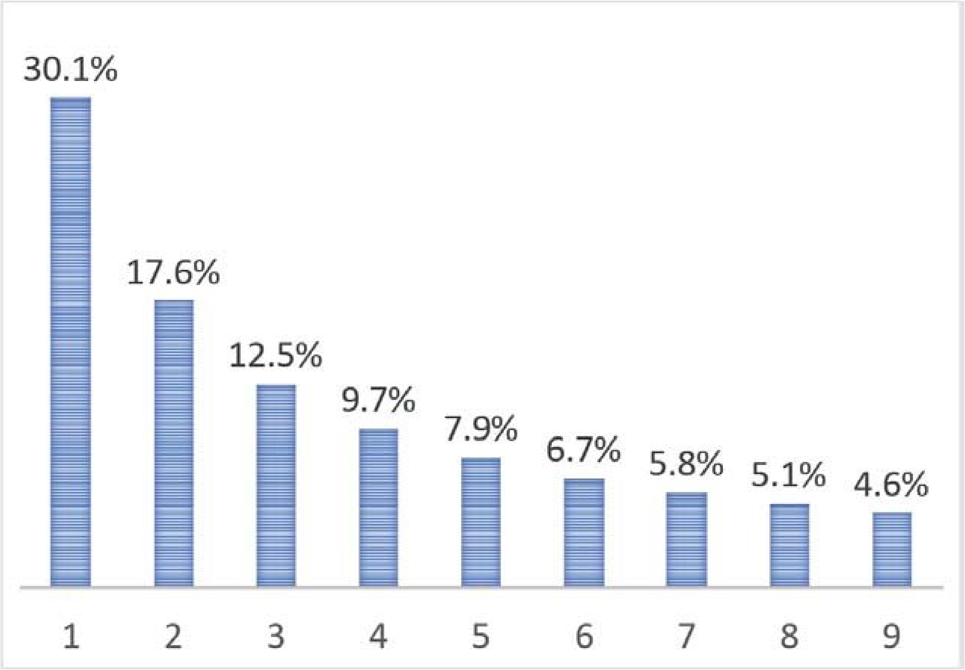

FIGURE 1.1: The Probability distribution of Benford's Law.

The distribution of leading digits according to Benford's Law shows a significant deviation from equal probability. The following graph show percentage of leading digits according to Benford's Law.The definition above, though straightforward, presents challenges when applied to real data sets. A key issue is that the values of $log_{10}\frac{d+1}{d}$ are irrational, while the number of times a first digit d appears in a data set must be an integer, leading to observed frequencies that are always rational.

One way to address this is to consider infinite sets, but this isn't feasible for many real-world situations, where data sets are finite, like the number of counties or trading days. Thus, while the original definition works well for mathematical studies of sequences and functions, it isn't practical for many real-world data sets. Therefore, the definition needs to be adjusted.

Definition 1.1.2 *A data set satisfies Benford's Law for the Leading Digit if the probability of observing a first digit of d is approximately* $log_{10}(\frac{1+d}{d})$.

The term "approximately" in the definition is unclear and needs further clarification. Defining it is difficult, especially with large data sets, where tests like chi-square can become too sensitive to minor variations. For now, we'll interpret "approximately" as a good visual match, which works well in many cases.

1.2 Benford's Set:

Benford did not specify which data sets should follow the expected frequencies, except for mentioning natural events and scientific phenomena.Additionally, the logarithm of the difference between the largest and smallest values should ideally be an integer (like 1, 2, 3, etc.). These conditions describe a perfect Benford set. However, for a reasonable fit to Benford's Law, the data only needs to roughly follow this geometric pattern.

The following are practical guidelines for determining whether a data set is likely to follow Benford's Law.

- Records should represent the sizes of facts or events, such as town and city populations, river flow rates, or the sizes of celestial bodies.
- The data should not have any built-in minimum or maximum values, except for a minimum of 0 for data that can only be positive numbers (like election results, population counts, or inventory counts). A minimum of 10 is acceptable if all records below 10 are removed to prevent small, irrelevant amounts from affecting the results
- Another important consideration is that there should be more small records than large ones in the data set. The average value should typically be lower than the median value, and the data should not be tightly clustered around a single average.

However, there are certain types of data that do not follow Benford's Law. Here are a few examples:

1. Assigned Numbers:
   - Telephone numbers: These are assigned based on a specific system and do not follow a naturally occurring distribution.
   - Social Security numbers: These are also assigned sequentially or based on specific rules rather than arising from a natural process.
2. Uniform Distributions:
   - Lottery numbers: These are randomly selected and each number has an equal chance of being chosen.
   - Randomly generated numbers: If a set of numbers is uniformly generated, each digit has an equal likelihood of appearing in any position.
3. Numbers with a Set Range:
   - Exam scores: Often fall within a limited range (e.g., 0-100) and are not distributed in a way that aligns with Benford's Law.
   - Human heights: Typically fall within a narrow range (e.g., 150-200 cm) and do not show the wide variance needed to match Benford's distribution.
4. Constrained Data Sets:
   - Product prices: Prices are often influenced by psychological pricing strategies, such as pricing items just below a round number (e.g., 9.99), which disrupts a natural distribution.

- Measurement data with specific units: Such as temperatures recorded in degrees Celsius within a limited range.

5. Synthetic or Designed Data:

   - Certain financial reports: If manipulated or fabricated, the numbers may not follow Benford's Law, which is sometimes used as a method for detecting fraud.

1.3 Example:

Let's look at a real-world example of Benford's Law. I downloaded the population data for top 51 countries in the world from the world bank data 2024. Do the leading digits of these population numbers match the expected distribution? After examine, this dataset is not theoretically perfect for testing Benford's Law because of the range of population number, when I test the data in Excel its shows some changes from Benford's Law,

Digit	Frequency	Observed Probability	Benford's probability
1	10	0.0.192308	0.301
2	5	0.096154	0.1760
3	14	0.269231	0.1250
4	8	0.153846	0.970
5	5	0.0.96154	0.790
6	4	0.0976923	0.670
7	1	0.019231	0.580
8	3	0.057692	0.510
9	1	0.019231	0.460
N	51		

TABLE 1.2: First Digit Benford's Law World's population Data

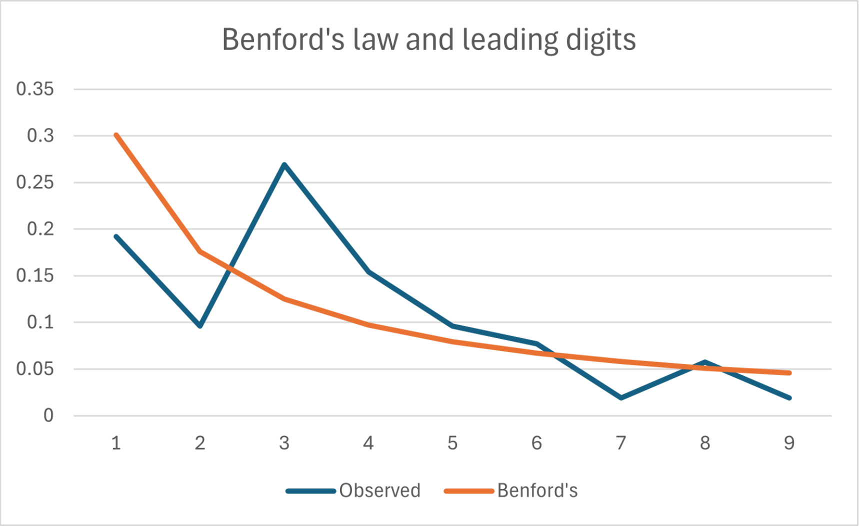

FIGURE 1.2: World population Data

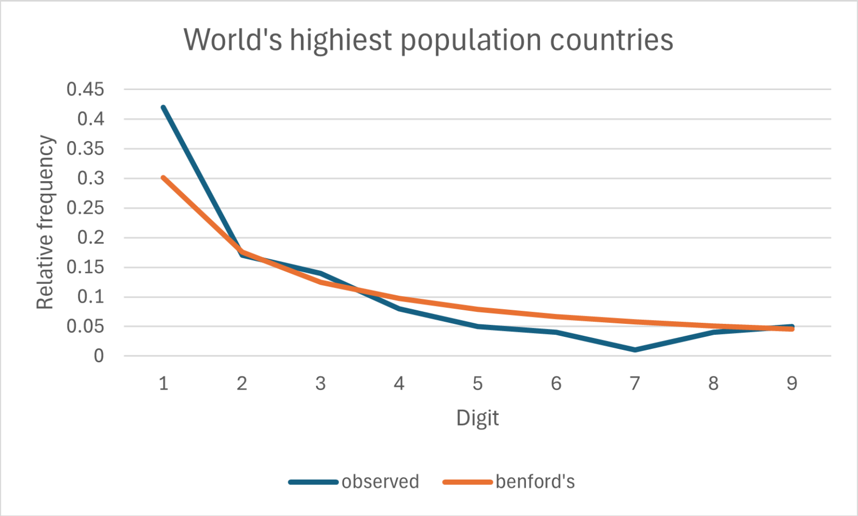

FIGURE 1.3: First Digit Test World population data

But if we increase the N value from 51 to 100 we get an improved Benford's graph line.

Digit	Frequency	Observed Probability	Benford's probability
1	42	0.0.807692	0.301
2	5	0.326923	0.1760
3	14	0.269231	0.1250
4	8	0.153846	0.970
5	5	0.0.96154	0.790
6	4	0.076923	0.670
7	1	0.019231	0.580
8	3	0.076923	0.510
9	1	0.096154	0.460
N	100		

TABLE 1.3: First Digit Benford's Law World population countries

The result became more accurate when we increase the N number from 100 to 150. For a data set to confirm well with Benford's Law, it usually needs at least 1,000 entries. If there are fewer than 1,000, the results might show larger deviations from the expected Benford proportions, but the tests can still be useful.Another rule is to avoid testing the first two digits in data sets with fewer than 300 entries. For smaller data sets, use the first digit test even though it's not perfect. If there are fewer than 300 entries, you can just sort them from largest to smallest and look through them for any unusual patterns.For example if we take the 3144 UK countries population the results foe Benford's law is more accurate,

1.4 Applications:

Before we discuses how Benford' Law works there is a question is this law used for practical purpose? Yes,auditors and analysts apply Benford's Law to examine

Digit	Frequency	Observed Probability	Benford's probability
1	946	30.08906	30.1
2	589	18.7341	17.60
3	382	12.15013	12.50
4	296	9.414758	9.70
5	241	7.665394	7.90
6	203	6.456743	6.70
7	163	5.184478	5.80
8	169	5.375318	5.10
9	155	4.930025	4.60
N	3144		

TABLE 1.4: UK countries Population

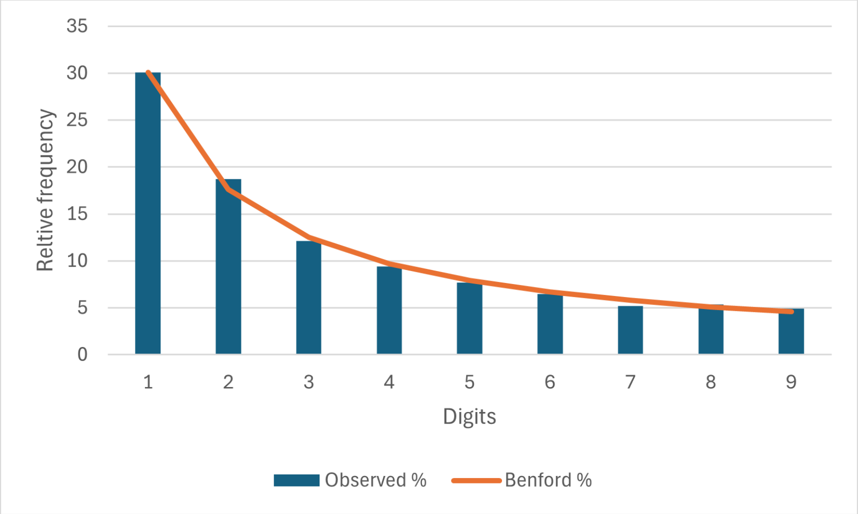

FIGURE 1.4: UK countries population

leading digit distributions in financial statements.Significant deviations from the expected pattern can indicate potential fraud.Tax authorities use Benford's Law to spot anomalies in tax data. Unnatural leading digit distributions may suggest manipulation of income or expenses.Organizations assess the integrity of applications and decision-making documents by comparing leading digit distributions to Benford's Law.

- Fraud Detection: One of the most famous applications of Benford's Law is in detecting financial fraud. Since many financial figures (like sales numbers, expenses, and incomes) in an organization naturally follow Benford's distribution, deviations from this expected distribution can indicate manipulation or falsification.

  Suppose a company reports its financial data to auditors, and the first digits of their expense reports are analyzed. Normally, you'd expect about 30% of the first digits to be '1'. However, the auditors notice that '7' appears more frequently than '1'. This deviation from Benford's Law might suggest that the numbers have been manipulated, possibly to meet certain financial targets or conceal losses.

- Data Validation: Benford's Law is also useful for validating data in fields such as scientific research, public data reporting, or accounting. If data naturally follows Benford's distribution, significant deviations could signal errors or inconsistencies.

  Consider a dataset containing population figures from various cities around the world. When applying Benford's Law, the expected distribution of first digits should roughly follow the law's prediction. If the first digit distribution significantly deviates from Benford's Law, it could indicate issues like data entry errors, inconsistencies in the way data is reported, or even deliberate alteration of data.

- Anomaly Detection: Benford's Law is applied in anomaly detection across various industries, such as cybersecurity, healthcare, and tax enforcement. When monitoring systems, any deviation from expected patterns (like those predicted by Benford's Law) can signal an anomaly that requires further investigation.

  In cybersecurity, Benford's Law can be applied to network traffic data. If network traffic volume is analyzed over time, and the distribution of the first digits of traffic counts suddenly shifts away from the expected Benford distribution, this could indicate unusual activity, such as a cybernetic or data breach.

- Election Data Analysis Benford's Law has been controversially applied to election data to detect potential fraud. The idea is that if vote counts for candidates follow the expected distribution, then the election process is likely legitimate; deviations might suggest manipulation.

  FOR example,During an election, analysts could examine the first digits of the vote counts from various districts. If, for example, the digit '1' appears less frequently than expected, and other digits, like '8' or '9', appear more frequently, this could raise suspicions that the vote counts have been tampered with or that there's an error in data reporting.

- Accounting and Auditing Auditors use Benford's Law to assess whether financial statements or other records have been manipulated. If an individual or company is fabricating data, they might not naturally produce figures that adhere to Benford's distribution.

  For example,an auditor might apply Benford's Law to the financial records of a company. If the digits of the revenue figures, expense reports, or tax filings deviate significantly from the expected distribution, it could prompt a deeper investigation into the accuracy of the financial records.

Chapter 2

Foundation and Properties

In large, naturally occurring datasets, the leading digits follow Benford's Law because of the logarithmic scale of growth. For this first identify and extract the leading digits from the dataset.Now,Calculate the frequency of each leading digit (1-9). The next step is Use Benford's Law formula to determine the expected frequency distribution

$$prob(d) = log_{10}(1 + \frac{1}{d})$$

The next stage of Benford's law is Compare the observed frequencies to the expected frequencies. When individuals manipulate numbers, they rarely consider the natural frequency of leading digits. This manipulation leads to an unnatural distribution, deviating from Benford's expected pattern.

Benford's Law is based on logarithms and is used to create a set of data that fits Benford's Law perfectly. It also helps to develop a general rule for significant digits. The key idea is that the mantissas of the logarithms of numbers should be evenly distributed.The mantissas is the fractional part of the right of decimal point,which range usually 1 to 9. The second part of characteristic of log, that's a left of decimal point. Now, we drive a formula for the expected frequencies of digits in the list of numbers. Let D_1 represent first digit,D_2 second digit,

$$prob(D_1 = d_1) = log(1 + \frac{1}{d_1}); \qquad d_1 \varepsilon \{1, 2, ..., 9\} \tag{2.1}$$

$$prob(D_2 = d_2) = \sum_{d_1=1}^{9} log(1 + \frac{1}{d_1 d_2}); \qquad d_2 \varepsilon \{0, 1, 2, ..., 9\} \tag{2.2}$$

The formula for first two digits D_1D_2 is,

$$prob(D_1D_2 = d_1d_2) = log(1 + \frac{1}{d_1 d_2}); \qquad d_1d_2 \varepsilon \{0, 1, 2, ..., 9\} \tag{2.3}$$

The general form of logarithm basis of Benford's law, which is used to calculate the first, first two, first three digits can be written as,

$$prob(D_1 = d_1, ..., D_k = d_k) = log[1 + (\frac{1}{\sum_{i=1}^{k} 10^{k-i}})]$$

Benford's Law predicts how often different digits appear as the first digit in a set of numbers. These proportions are irrational numbers, meaning they can't be expressed as simple fractions. This doesn't make them unreasonable or unstable, just mathematically complex.

As the number of records increases, the proportions of first digits in the data will get closer to Benford's predicted values. Benford's Law works best with large numbers and assumes each number has many digits.

To match Benford's Law well, numbers should generally have at least four digits. If the data includes numbers with fewer digits, there might be a slight bias toward lower digits, but this bias isn't significant if there aren't too many two- or three-digit numbers mixed in. For example, in census data with 19,000 numbers, about 1,000 are less than 100, but the overall fit to Benford's Law remains impressive.[9]

2.1 Pinkham's scale invariance theorem

Benford law goes through many modernization after the publishing of Benford's paper.There are many theorems that are relevant to using Benford's Law in data analysis, but the most important theorem is scale in-variance theorem.

Pinkham (1961) proposed that any law governing digital distributions should be scale-invariant. For instance, if the areas of the world's islands or the lengths of its rivers followed a certain law, it should not matter whether these measurements were in miles or kilometers. Pinkham demonstrated that Benford's Law is scale-invariant under multiplication, meaning that if you multiply all numbers in a data set conforming to Benford's Law by a nonzero constant, the resulting set will also follow Benford's Law.

He further showed that only Benford's Law maintains its digit frequencies under multiplication. Therefore, if a list of numbers has digit frequencies different from Benford's Law, multiplying by a constant will change these frequencies. It stands to reason that the closer a set conforms to Benford's Law before multiplication, the closer it will conform afterward.Pinkham noted that a sampling experiment on Benford's Law would be fascinating and much easier to conduct today compared to fifty years ago.Pinkham's proof is based on the fact that only a cumulative distribution function (CDF) of mantissas.

Theorem 2.1.1 *If the numbers $x_1, x_2, x_3, ..., x_N$ in a data field conform to Benford's Law, any new field formed by multiplying the xi values by a nonzero constant c will also conform to Benford's Law.*

To explain this theorem Let's explain the example of the tallest buildings in the world and how it relates to Benford's Law and Pinkham's Scale In-variance Theorem in simple terms.We analyzed the heights of the 58 tallest buildings in the world to see if they follow Benford's Law.Suppose we convert all building heights from meters to another unit by multiplying by a constant (e.g., feet, where 1meter=3.281 feet). According to Pinkham's theorem, if the heights in meters follow Benford's Law, the converted heights in feet should also follow Benford's Law. For example, the height 828 meters has a leading digit of 8, and 2716 feet has a leading digit of 2. For each height, we looked at the first digit. For example, the height 828 meters has a leading digit of 8, and 2716 feet has a leading digit of 2.We counted frequencies how many times each digit (1 through 9) appears as the first digit in meters and feet.Then compared our counts to the expected distribution according to Benford's Law.

Although the observed frequencies did not perfectly match the theoretical values, they were close enough to confirm the law's applicability because this helps us understand that Benford's Law is not dependent on the unit of measurement. So, whether we look at the heights in meters or feet, the leading digits should follow the same pattern. Although it does not perfectly match up with the Benford's Law

Digit	Observed Prob(m)	Observed Prob(ft)	Benford's prob
1	41.3793	25.8620	30.1
2	20.6896	12.0689	17.60
3	6.8965	10.3448	12.50
4	13.793	12.0689	9.70
5	1.7241	15.5172	7.90
6	8.6206	6.8965	6.70
7	1.7241	5.1724	5.80
8	5.1724	10.3448	5.10
9	0	4.930025	1.7241
N	58		

TABLE 2.1: First Digit Benford Test Tallest Building Data in Meter and Feet

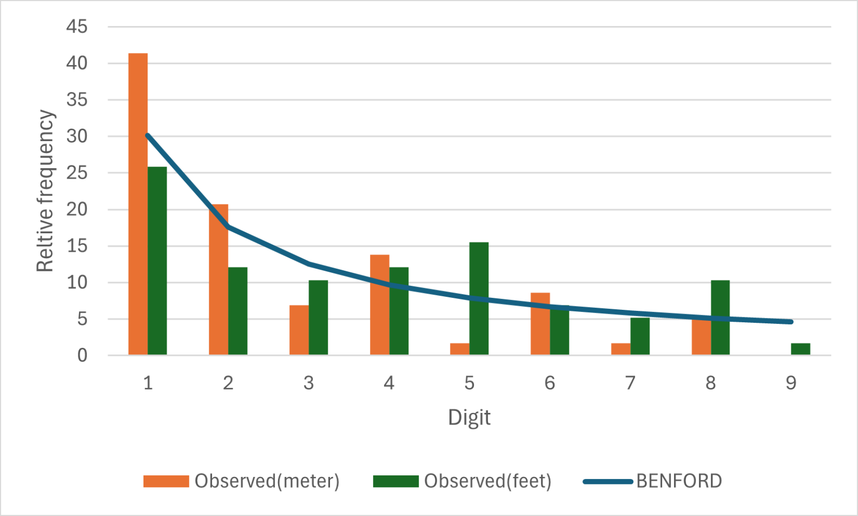

FIGURE 2.1: Invariance Theorem for Tallest Building Data

graph, but it is reasonably a good fit.Maybe with few tweaking of the geometric function, we can model Benford's Law with better accuracy.

Also this law is useful for checking if data is manipulated. If a dataset follows Benford's Law in one unit, it should also follow it in another unit. If not, it might be a sign of fraud or manipulation.

2.2 Benford Properties for Sequences:

Many naturally occurring sequences in mathematics, particularly those that grow exponentially or cover a wide range of magnitudes, tend to follow Benford's Law. This includes geometric sequences, the Fibonacci sequence, factorials, and sequences of powers. However, not all sequences especially those with linear growth like arithmetic sequences, will conform to Benford's Law. Understanding the growth pattern and the distribution of leading digits in these sequences helps explain why certain sequences follow this law while others do not. This section aims to present three essential properties that indicate a sequence of constants follows Benford's Law:

- The fractional parts of its decimal logarithm are uniformly distributed between 0 and 1.
- The distribution of its significant digits remains unchanged when the scale is altered.
- The distribution of its significant digits is continuous and remains consistent even when the numerical base is changed.

Understanding the growth pattern and the distribution of leading digits in these sequences helps explain why certain sequences follow this law while others do not. A geometric sequence g_n can be represented as:

$$a, ar, ar^2, ar^3, ..., ar^n - 1$$

They grow exponentially and span several orders of magnitude, which helps them conform to Benford's Law.Consider a geometric sequence with a starting number $a = 1$ and a common ratio $r =$ 3:1,3,9,27,81,243,729,. . . .

In Fibonacci Sequence f_b each term is the sum of the two preceding ones, starting from 0 and 1.For example:0,1,1,2,3,5,8,13,21,. . . . This is because despite not being strictly geometric, the ratio between successive Fibonacci numbers approximates the golden ratio (1.618), leading to a logarithmic-like distribution.

Factorials f_c grow extremely rapidly, covering many orders of magnitude quickly, which tends to produce a distribution of leading digits similar to Benford's Law.The product of all positive integers up to a given number. Example $1!, 2!, 3!, 4!, \ldots$

In arithmetic Sequence a_n each term is obtained by adding a constant difference to the previous term, a,a+d,a+2d,a+3d,. . . . They do not span several orders of magnitude rapidly enough and often do not produce a logarithmic-like distribution.

.

g_n	a_n	f_b	f_c	2^n	Benford's prob
30	19.5	30	27	29.5	30.1
17.5	18	17.5	14.5	18.5	17.60
12.5	18	12.5	11	12	12.50
9.5	19.5	9	6	10	9.70
8	18	8.5	6	8	7.90
6.5	1.5	6	5	6.5	6.70
5.5	2.5	5.5	3	5.5	5.80
5.5	1.5	6	7	5.5	5.10
5	1.5	4.5	5.5	4.5	1.7241

TABLE 2.2: Benford's Law in different sequences

The table and chart provided distribution of leading digits in these mathematical sequences compared to the expected distribution according to Benford's Law

Benford's Law is particularly applicable to sequences that grow exponentially (like geometric and Fibonacci sequences) because they span many orders of magnitude. Linear sequences, like arithmetic sequences, do not conform to Benford's Law as closely because of their normal growth.Besides of these general sequences,Most, but not all, exponentially increasing sequences tend to follow Benford's Law. If a sequence does follow Benford's Law from one starting point, it will follow it from

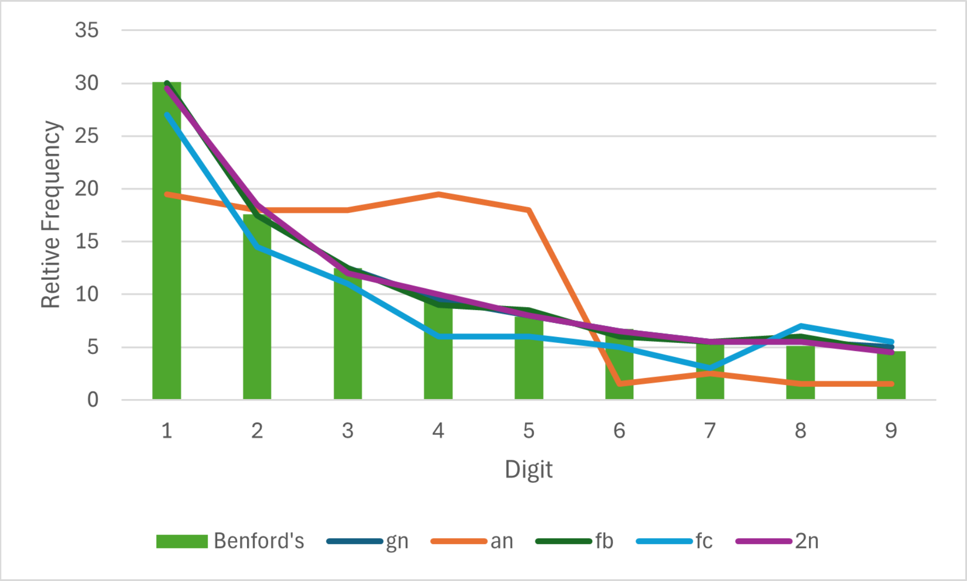

FIGURE 2.2: Benford's law and sequence

any starting point.But Sequences that increase or decrease in a polynomial manner, or their reciprocals, do not follow Benford's Law. Sequences that grow or shrink at a rate faster than exponential (super-exponentially) generally follow Benford's Law, though not in every case, depending on the starting point.

Chapter 3

Measuring Benfordness

3.1 Primary Benford's Law Tests:

Benford's Law for the second digit is an extension of the first digit law, providing another tool for analyzing the distribution of digits in naturally occurring datasets. It is particularly useful when the first-digit test isn't conclusive, allowing deeper analysis of the data's authenticity or integrity. The second digit test is good for spotting biases in data. Biases happen when people aim for specific numbers to get around control limits. This also occurs in marketing when stores set prices ending in 9, like in supermarkets or discount chains. Although, The second digit test is also a high-level test, but it can be useful for detecting biases in the data or rounding-up behavior by corporate controllers.

Benford's Law for the second digit states that in many naturally occurring datasets, the second digits are not uniformly distributed but instead follow a logarithmic distribution. This means that certain digits appear more frequently as the second digit than others. To perform the second digit Benford test in Excel on UK data and include both counts and expected frequencies.We use *MAD* command on an adjacent column to extract the second digit.Then count how often each digit appears as the second digit *Actual* proportion in Table 3.1. Use Benford's Law to determine the expected frequency of each second digit as$P(D$.The Benford's proportion for second digit can be calculate by multiply probability of each digit with 'N'.

Digit	Actual	P(D)	Benford
0	370	0.1197	376.3368
1	340	0.1139	358.1016
2	335	0.1088	342.0672
3	320	0.1043	327.9192
4	307	0.1003	315.3432
5	325	0.0967	304.0248
6	331	0.0934	293.6496
7	297	0.0904	284.2176
8	287	0.0876	275.4144
9	231	0.085	267.24

TABLE 3.1: Second Digit UK Population Data

The graph in 3.1 show comparison between actual vs. expected frequencies of the second digits.The graph indicates that the dataset shows a general conformity to Benford's Law for most of digits, especially for the digit '1'.

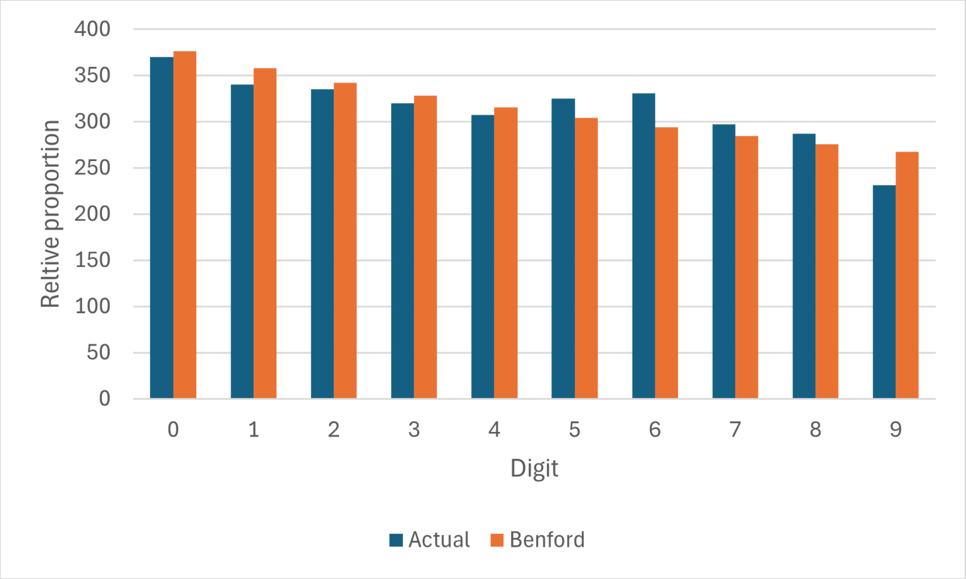

FIGURE 3.1: Second Digit test

The second-digit proportions are based on the assumption that the first digits follow Benford's Law. If the first digits have more lower digits than expected, the second-digit proportions will be more skewed than those. Conversely, if the first digits have more higher digits, the second-digit proportions will be closer to a uniform distribution.

The first-two digits test is more precise than the first digits test and helps identify unusual digit repetitions and potential biases in the data. This test combines information from both the first and second digit graphs into a single, more detailed graph. Although you could create separate first and second digit graphs using the same data, the first-two digits graph provides more comprehensive insights. To perform a Fist two-digit Benford's Law test on UK population data set in Excel, we use an command $IF(A1 > 9, LEFT(A2, 2)$ to extract the first two digits.Create a list of all possible two-digit combinations from 10 to 99 in a column which represent as $Actual$ proportion.

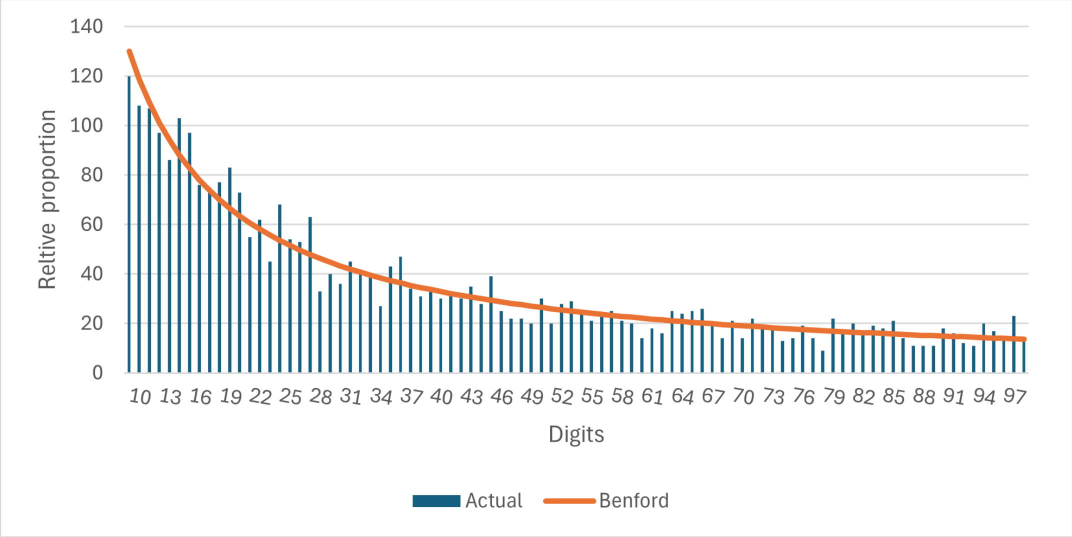

FIGURE 3.2: First Two Digit Test for UK population data

Next, we calculate the expected frequency for each two-digit combination using Benford's Law formula 2.3 in chapter 2.The Benford's proportion for First-two digit can be calculate by multiply probability of each digit with total number of entries.The graph in B.1 show comparison between actual vs. expected frequencies of the second digits.The graph indicates that the dataset shows a general conformity to Benford's Law for most of digits, especially for the digit '1'.

The first-two digits test provides more detailed insights than the separate first and second digit tests and usually results in smaller audit samples. It is recommended as the main Benford's Law test for most cases, particularly with larger datasets, though there are exceptions for very small datasets. This test is also useful for identifying biases in the data.

3.2 Z-Statistics:

To determine if a dataset conforms to Benford's Law, we often rely on both professional judgment and statistical methods. In some cases, it's straight forward to spot deviations from expected proportions by simply looking at the data, as certain digits will stand out more than others. However, when these deviations aren't obvious, a more empirical approach is needed. In such situations, examiners should use statistical techniques to identify which digits show significant discrepancies from the expected patterns.[14]

Definition 3.2.1 *The Z-statistic is used to evaluate how well a dataset conform with Benford's Law. To calculate the Z-statistic, you use the following formula,*

$$Z = \frac{|AP - EP| - (\frac{1}{2N})}{\sqrt{\frac{EP(1-EP)}{N}}} \tag{3.1}$$

The term $(\frac{1}{2N})$ is a continuity correction factor used only if it is smaller than the primary term in the numerator. This method helps us assess whether our data follows Benford's Law or if significant deviations suggest something unusual in the source of the data. When applying the Z-statistic,a examiners can choose a significance level, such as 5%, to determine the acceptable amount of variance. If the Z-statistic for a digit is within this range, the observed frequency of that digit in your data is close enough to what Benford's Law predicts that you would conclude there's no significant deviation. This suggests your data likely conforms to Benford's Law.If the Z-statistic falls outside this range(e.g., < -1.96 or > 1.96), it indicates a significant deviation. The observed frequency of that digit is either much higher or lower than what Benford's Law predicts. This suggests that your data may not follow Benford's Law.

Let's say running a test on the first digit again the dataset of Example1 world population countries. Here's an example of the detailed results from a made-up dataset. The "Expected Proportion" in this example follows Benford's Law. If we focus on digits that are statistically significant at the 95% level, numbers 1 and 7 will be scrutinized since they don't align with Benford's Law. This is because the Z-statistics for these digits exceed the threshold of 1.96.

If we focus on digits that are statistically significant at the 95% level, numbers 1 and 7 will be scrutinized since they don't align with Benford's Law. This is because the Z-statistics for these digits exceed the threshold of 1.96.

In general,the expected proportion appears twice in the denominator. The influence of the expected proportion is such that, for any given difference, a smaller

Digit	Count	Ap	Ep	z-stat
1	42	0.42	0.301	2.309756561
2	17	0.17	0.176	0.026621743
3	14	0.14	0.125	0.288195209
4	8	0.08	0.097	0.442325868
5	5	0.05	0.079	1.101195523
6	4	0.04	0.067	1.122682799
7	1	0.01	0.058	4.321662606
8	4	0.04	0.051	0.306186218
9	5	0.05	0.046	0.183532587
N	100			

TABLE 3.2: z-statistic Benford's Test

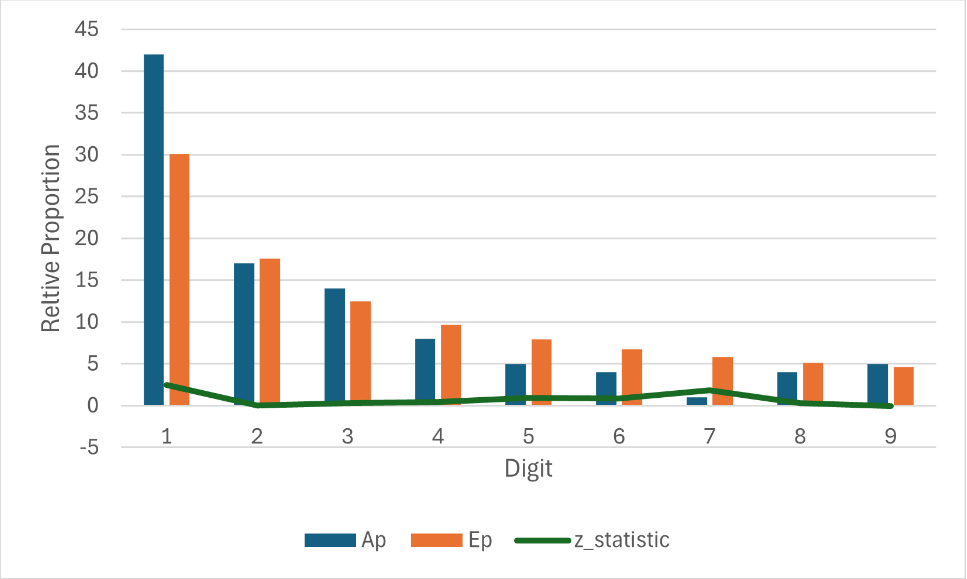

FIGURE 3.3: Z statistic for world population data

expected proportion results in a larger Z-statistic.A 0.19 percent difference is more significant when the expected proportion is lower, meaning differences in higher digits (with lower expected proportions) are more significant than those in lower digits.

The continuity correction in the Z-statistic formula has minimal impact, but as the data set size increases, the Z-statistic for any deviation also grows. This means a small deviation may be insignificant in a small dataset but significant in a larger one. For example, a dataset of 100,000 records could yield a Z-statistic of 5.178, with the Z-statistic increasing proportionally less than the dataset size due to the square root in the formula.

The Z-statistic has an issue with excess power, meaning that as a dataset increases in size, even smaller deviations become significant.In other words,When using the Z-statistic, in large datasets, the Z-statistic may only allow very small deviations, potentially leading to the unjustified rejection of the dataset. For example, if we take a very big data set, dataset of 55000 units instead of 500, the proportions for all digits may remain the same, but the Z-statistic would now flag even smaller differences as significant.

For example,if we examine a data set of Example 4 of stock exchange prices, where N=5550,The digits that are now statistically significant include 1, 2, 3, 6, 7,8 and 9. This might mean the data doesn't conform to Benford's Law, making the analysis questionable, or it could be due to the excess power problem. A possible solution is to disregard the exact Z-statistic values.

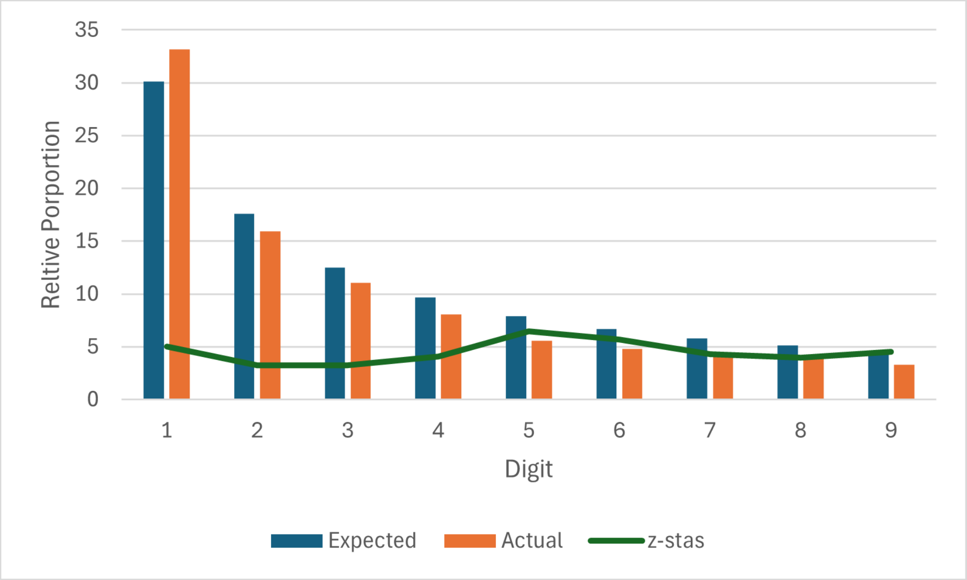

FIGURE 3.4: Z statistic for Stock Exchange Data

where z-stat for each digit is greater then the significant. So,while using z-statistic test , be ready about the excess problem problem; some time a set is so large the z-statistic only marge minor fluctuations and dataset might be inappropriately rejected.

Chapter 4

Hypothesis Testing:

Observing deviations from Benford's Law can be a valuable tool for detecting anomalies in data. To effectively measure these deviations, we turn to the concept of statistical hypothesis testing. This approach allows us to assess and validate assumptions about a random process based on a limited set of observations. In the context of our discussion, the random process involves the population from which our numerical data originates. The hypotheses under consideration are as follows:

Null hypothesis;

H_o=The digits come from dataset conform Benford's Law,

Alternative Hypothesis;

H_1=The digit come from data set does not obey Benford's Law

Based on the outcome of the test, we either accept the null hypothesis or reject it in favor of the alternative hypothesis. If we reject the null hypothesis, it may indicate that the data warrants closer examination, as deviations from Benford's Law could suggest potential issues or irregularities. An important aspect of this process is the significance level, often referred to as the p-value. In statistical analysis, the results are not absolute; instead, our conclusions are made with a certain degree of confidence, typically less than 100%. While the exact definition of the p-value is complex, it can be understood as a measure of how extreme the observed test statistic is, assuming that the null hypothesis is true.The process of conducting a statistical test follows this general workflow:

- Collection of data: Gather your dataset and extract the first digits of each number.Count frequencies Count how many times each digit (1 through 9) appears as the first digit in your dataset.

- Expected Count: Benford's Law gives the expected probability for each digit as the first digit by using formula,

  $$prob(d) = log_{10}(1 + \frac{1}{d})$$

  Calculate expected frequency of each digit 1 to 9 and then multiply each probability by the total number of observations to get the expected frequency for each digit. The next step will be a comparison between expected and actual count that can be generated by help of different type of statistic testing like $chi - square$

- Results: If the calculated statistic exceeds the critical value, the null hypothesis is rejected at the chosen significance level. Conversely, if the statistic is below the critical value, the null hypothesis is accepted.

For hypothesis testing related to Benford's Law, two widely used methods are Pearson's Chi-Square goodness-of-fit test and the Cho-Gaines d statistic. Let's apply these tests to our four example datasets.[2]

4.1 Chi Square Test:

Z-statistics for individual digits can't be combined to measure overall nonconformity. To get a broader assessment, you can use tests like chi-square or Kolmogorov-Smirnov, which evaluate all digits together to see if the data follows Benford's Law. The maximum likelihood test statistic is typically not used on its own. Instead, Pearson's chi-square test statistic is often employed as an approximation. While the results are similar for large samples, the maximum likelihood statistic tends to produce more accurate results for smaller samples.[4]

Definition 4.1.1 *A chi-square is calculated by formula,*

$$chi - square = \sum_{i=1}^{k}(\frac{(AC - EC)^2}{EC})$$

where AC and EC represent the actual count and expected count respectively, And K is number of "counts", in case of Benford's law K=9.

It is important to notice that if the actual term is far from expected term the leading sum is very big and if they are quite closer the square term would be small. Deviations from Benford's Law can point to unexpected in data. To measure these deviations, we use statistical hypothesis testing discussed in start of chapter.

For sufficiently large values of n, Pearson's chi-square test statistic approximates a chi-square distribution with $k - 1$ degrees of freedom, denoted as χ^2_k.To perform a chi-square test for Benford's Law, follow these steps after using hypothesis testing to calculate the AC and EC to determine if your observed data conforms to the expected distribution predicted by Benford's Law.

- Calculate Chi-Square Statistic: After computing actual and expected count for each digit, compute the squared difference between observed and expected frequencies, divide by the expected frequency, and then sum these values by using formula
$$chi - square = \sum_{i=1}^{k}(\frac{(AC - EC)^2}{EC})$$
- Determine the p-Value Use the chi-square statistic and the degrees of freedom df to find the p value.Degrees of freedom df_{k-1}, where k is the number of categories (digits 1 through 9, so $k = 9$, and $df = 8$.
- Interpret the Result: Compare the p-value with your significance level (e.g., 0.05). If the p-value is less than the significance level, reject the null hypothesis that the data follows Benford's Law.[6]

Now, to see how chi-square works lets consider the example of world population countries data set.As N=100,or each digit from 1 to 9 listed in column , we calculate the observed count of data values with that digit as the first significant digit. For example, the formula in next cell is =COUNTIF($B2$:$B100$,Q2), which counts occurrences of the digit in the data range.

Digit	P(d)	EC	AC	Chi Square
1	0.301	30.1	42	4.704651
2	0.176	17.6	17	0.020455
3	0.125	12.5	14	0.18
4	0.097	9.7	8	0.297938
5	0.079	7.9	5	1.064557
6	0.067	6.7	4	1.08806
7	0.058	5.8	1	3.972414
8	0.051	5.1	4	0.237255
9	0.046	4.6	5	0.034738
N	100		df	11.60011
			p	0.169957

TABLE 4.1: Chi-Square world population data

The expected count for each digit, based on Benford's Law, is shown in next column. For instance, cell S2 contains the formula =S11*LOG10(1+1/N2), where Q11 calculates the total count with =SUM(Q2:Q10).

To determine if the data conforms to Benford's Law, we compute the p-value in cell using the formula =CHISQ.TEST(U2:U10,T2:T10). Since the p-value in 4.1 of 0.169957 is greater than the significance level of 0.05, we conclude that there is insufficient evidence to reject the null hypothesis that the data follows Benford's Law.

If we try ch-square test for bigger data set that is UK population Data,where N=3144 the result are shown in 4.2.Where p=0.576866, again greater then significance level.

4.2 Kolmogorov-Smirnov Test:

The Kolmogorov-Smirnov (K-S) test can also be used to test if a dataset follows Benford's Law. Unlike the chi-square test, which compares observed frequencies to expected frequencies, the K-S test compares the cumulative distribution of the observed data with the cumulative distribution expected under Benford's Law. In the context of Benford's Law, it calculates the cumulative probabilities for each digit and compares them to the actual observed cumulative values in a dataset. The K-S

Digit	AC	P(d)	EC	Chi Square
1	946	0.301	946.34	0.000125
2	589	0.176	553.344	2.297577
3	382	0.125	393	0.307888
4	296	0.097	304.96	0.263716
5	241	0.079	248.37	0.219044
6	203	0.067	210.648	0.277676
7	163	0.058	182.352	2.05372
8	169	0.051	160.344	0.467285
9	155	0.046	144.624	0.744423
N	3144		df	6.631454
			p	0.576866

TABLE 4.2: chi Square UK population Data

test focuses on the largest difference between these cumulative distributions to assess whether the data significantly deviates from the expected pattern. The K-S test is generally applied to data that is expected to conform closely to Benford's Law, aiming to detect minor deviations.[8]

Null Hypothesis (H_0): The data follows Benford's Law closely, meaning the Kolmogorov-Smirnov statistic will be less than the critical value, indicating high conformity.

Alternate Hypothesis $(H1)$: The data does not conform to Benford's Law, meaning the Kolmogorov-Smirnov statistic will exceed the critical value, signaling deviations from Benford's Law.

Here's how you can perform the Kolmogorov-Smirnov test for Benford's Law:

- Calculate the Observed Cumulative Distribution: Count Frequencies: Count the frequency of each first digit (1 through 9) in your dataset. Convert these frequencies into cumulative frequencies. For digit 1, the cumulative frequency is just the frequency of 1. For digit 2, it's the sum of the frequencies of 1 and 2, and so on.

  Calculate the Observed Cumulative Distribution: Divide each cumulative frequency by the total number of observations to get the observed cumulative distribution.

- Calculate the Expected Cumulative Distribution: By applying Benford's Law The expected probability for each digit d is given by equation,

  $$P(D) = log_{10}(1 + \frac{1}{d})$$

  Calculate the expected cumulative distribution by summing the probabilities up to each digit and calculate the cumulative sum of these expected probabilities.

- Calculate the Kolmogorov-Smirnov Statistic: Compute the Differences: Calculate the absolute difference between the observed cumulative distribution and the expected cumulative distribution for each digit.

  $$D = max|F_{obs}(d) - F_{exp}(d)|$$

Digit	Frequency	F$_o$	F$_e$	$\lvert F_o - F_e \rvert$
1	42	0.42	0.301	0.119
2	17	0.17	0.176	0.113
3	14	0.14	0.125	0.128
4	8	0.08	0.097	0.111
5	5	0.05	0.079	0.082
6	4	0.04	0.067	0.055
7	1	0.01	0.058	0.007
8	4	0.04	0.051	0.007
9	5	0.05	0.046	0.004
			D	0.128
			D_{crit}	0.136

TABLE 4.3: K-S test World population Data

i.e the maximum difference across all digits is the K-S statistic.

- Determine the p-Value Compare with Critical Value: The K-S test statistic can be compared against critical values from a Kolmogorov-Smirnov distribution table, or the p-value can be calculated based on the statistic and the number of observations. The formula for the critical value for the K-S test is,

$$\text{Kolmogorov-Smirnoff critical value} = \frac{1.36}{\sqrt{N}}$$

  where 1.36 is the constant for a significance level of 0.05 and N is the number of records.

- Interpret the Result: If the K-S statistic is larger than the critical value, or the p-value is smaller than the chosen significance level (e.g., 0.05), then there is sufficient evidence to reject the null hypothesis that the data follows Benford's Law.

After examine k-s test the for World Population Data, the data seems to follow Benford's Law. In Table 4.3 $D = 0.128$ and $D_{crit} = 0.136$.As $D < D_{crit}$ which means that we don't have sufficient evidence to reject null hypothesis.

As the Kolmogorov-Smirnov (K-S) test is used to compare expected and actual cumulative distributions, and it is especially sensitive when dealing with large

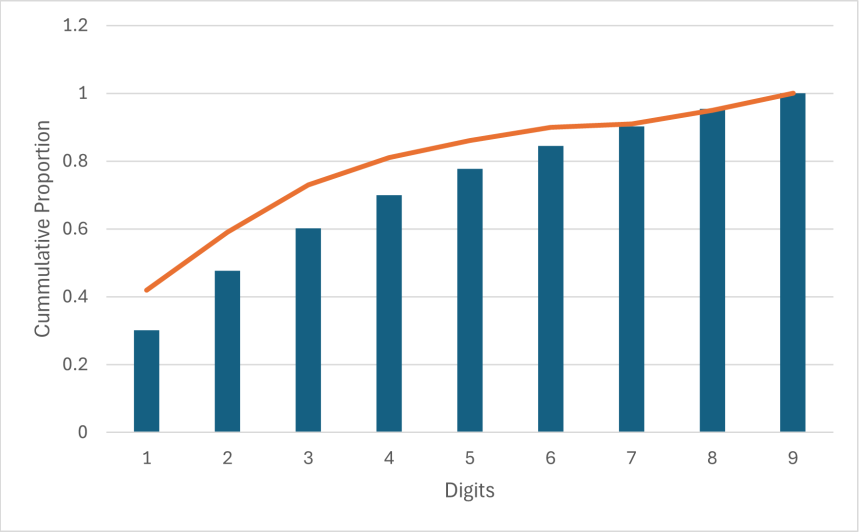

FIGURE 4.1: Cumulative density and actual proportion of World population Data

datasets. In the example, the expected and actual distributions are nearly identical, with only minor differences in the middle range. Since the largest observed difference is within the acceptable limit, the data is considered to follow Benford's Law. However, as the sample size increases, the K-S test becomes stricter, allowing less tolerance for deviations. This can make the K-S test challenging to apply in real-world scenarios with large datasets, where even small differences might lead to rejecting the data's conformity to Benford's Law.

4.3 Mean Absolute Deviation Test:

Previous tests incorporated the number of records to determine the critical value of the test statistic, leading to increased sensitivity to deviations as sample size grew. For datasets with 25,000 records or more, these tests required almost perfect conformity, which can be impractical for real-world data. To overcome this limitation, a test that does not depend on the sample size is necessary. The Mean Absolute Deviation (MAD) test addresses this need by calculating the average difference between actual and expected proportions across all bins, providing a clear measure of deviation independent of sample size.

Definition 4.3.1 *The formula for MAD is*

$$MAD = \frac{\sum_{i=1}^{k} |Ap - Ep|}{K}$$

Where k is number of bins(for first digit test k=9) AP is actual proportion and Ep represent expected proportion.

Additionally, the Mean Absolute Percentage Error (MAPE) is a related metric used in time-series analysis to gauge forecast accuracy. A lower MAPE indicates that predicted values closely match actual values, signifying reliable forecasts.The MAD is particularly useful because it measures accuracy in the same units as the data (proportions) and does not factor in the number of records, making it a more practical and straightforward tool for real-world data analysis.[13]

The Mean Absolute Deviation (MAD) involves three main steps:

- Deviation Calculation:The numerator of the MAD formula measures how much each actual proportion deviates from the expected proportion. Taking the absolute value ensures that the deviation is always positive.
- Summation and Averaging: The numerator aggregates all the absolute deviations across the bins. The total is then divided by the number of bins to get the mean absolute deviation.
- Interpretation: A MAD of 0.0006 represents the average difference between the observed and expected proportions. Higher MAD values indicate larger deviations. In comparative analyses, the MAD values reflecting greater deviations in the former.

While the MAD provides an indication of deviation, it lacks standardized critical values for assessing conformity. Guidelines from Drake and Nigrini (2000) offer practical benchmarks,summarizes MAD results from analyzing which shows that for first digit test $\alpha > 0.015$,

Close conformity — 0.000 to 0.004
Acceptable conformity — 0.004 to 0.008
Marginally acceptable conformity — 0.008 to 0.012
Nonconformity — greater than 0.012[5]

The Table 4.4 below provide the results of a MAD test for Benford's Law on a world population Data example.

Digit	Ap	Ep	∣Ap-Ep∣
1	42	30.1	11.9
2	17	17.6	0.6
3	14	12.5	1.5
4	8	9.7	1.7
5	5	7.9	2.9
6	4	6.7	2.7
7	1	5.8	4.8
8	4	51	1.1
9	5	4.6	0.4
		$\sum\lvert Ap - Ep\rvert =$	27.6

TABLE 4.4: MAD Worlds population Data

As the $\sum |Ap - Ep| = 27.6$, therefor for k=9

$$Mad = \sum |Ap - Ep|/k = 27.6/9 = 3.067$$

The overall MAD score of 3.0667>0.012 suggests that the dataset deviates significantly from Benford's Law.

4.4 Perfect Simulated Benford's law:

To assess the effectiveness of goodness-of-fit tests, it's important to validate these tests using a dataset that perfectly obey to Benford's Law. By starting with such a dataset, we can then deliberately introduce certain deviations (for example, by adding repeated numbers) to observe how these changes impact the test results. The mantissas of numbers that follow Benford's Law are evenly spread across the [0,1) interval. Therefore, one approach to generating a Benford-compliant dataset is to first create a set of mantissas uniformly distributed within this range, and then derive the corresponding numbers. This method forms the basis for constructing a synthetic dataset that follows Benford's distribution.

First step to create a perfect Benford's set, First, we need to calculate d, which is the difference between the logarithms of the upper and lower bounds. This difference must be an integer (1, 2, 3, etc.) for the results to form a Benford Set. The calculation is shown as:

$$d = log(ub) - log(lb)$$

The next step is to calculate r, the common ratio needed to generate a geometric sequence with exactly having N terms, starting with upper bound and ending at lower bound. This is done using the formula in Equation,

$$r = n\sqrt{10^d}$$

The simulated Benford Set closely follows the expected characteristics, ending just below upper bound. It forms a nearly perfect geometric sequence, with minimal deviations even with N=25,000 records. The sequence conforms well to Benford's Law, as shown by the chi-square test and the Mean Absolute Deviation (MAD). Overall, the set is highly accurate for a dataset of this size.

The general rule is that as N increases, the sequence becomes increasingly accurate. However, in this case, because the common ratio r is raised to a power that is a rational number, the sequence becomes re entrant. This means that it actually consists of four repeating sequences, each with 6,250 numbers. These numbers are identical digits,suppose to different d only by the placement of the decimal point, and they also share the same mantissas.

matlab-prettifier

```
 clc; close all; clear all;
N=443;
% Numbs=round( 4.^abs(3*(randn([1,N]))) ); % normal, base 10
Numbs=round( pi.^abs((1+pi)*(rand([1,N]))) ); % uniform, base pi

 if sum(isinf(Numbs)) >0
     beep
     disp('Overflow')
```

```
end

nodig= fix( log10(Numbs) );
NumbsF = fix( Numbs ./ (10.^nodig) );
h=histogram(NumbsF,9);
p=h.Values;
p=p/sum(p);
figure;

for d=1:9
    pBenford(d)=log10((d+1)/d);
end

plot(p,'r*')
hold on
plot(pBenford,'b')
legend('Data','Benford''s Law')

chisq=sum( ((p-pBenford).^2) ./ pBenford )*N;
figure
xx=linspace(0,60);
fpdf=chi2pdf(xx,8);
plot(xx,fpdf,'b');
hold on

a=max(fpdf);
plot([chisq chisq],[0 a],'k')
```

This code generates a set of random numbers, extracts their leading digits, and compares the distribution of these leading digits to the expected distribution according to Benford's Law. It also calculates and visualizes the chi-square statistic to test how well the data conforms to Benford's Law.

4.5 Conclusion:

This thesis has explored the principles and applications of Benford's Law, demonstrating its relevance in various domains such as population data analysis like population data and financial audits as stock exchange prices. By conducting statistical tests like the Chi-Square test, Kolmogorov-Smirnov test, and the Mean Absolute Deviation test on these examples, and the conformity of different datasets to Benford's Law was assessed, revealing the law's predictive power in identifying anomalies. The results confirm that Benford's Law is a powerful tool for detecting irregularities in large datasets, with the potential to uncover underlying patterns that may indicate manipulation or errors.

Appendix A

Examples Used in Project

A.1 Example 1:

set of world countries population Data This data set include top 100 most population countries,I download from Wikipedia site $Listofcountriesbypopulation(UnitedNations)$ visit Please find the code attached her

A.2 Example 2

length of largest Buildings of the world This data set include 58 logiest buildings of the world, please find the code attached

A.3 Example 3

UK Counteries Popullation Data This data set include 2023 population of 3144 countries of United Kingdom,

A.4 Example 4

Stock Exchange Price This data set include stock prices of 5550 companies downloaded by the

Appendix B

Examples Used in Project

B.1 Logarithm properties in Benford's Law:

Log plays the most important part in Benford's Law. The common logarithmic function, the base of the logarithmic function is 10 have following properties in general

- If a, m and n are positive integers and $a \neq 1$, then;

$$log - 10(mn) = log_{10}m + log_{10}n$$

- If m, n and a are positive integers and $a \neq 1$, then;

$$log_{10}(m/n) = log_{10}m\text{˘}log_{10}n$$

- If a and m are positive numbers, $a \neq 1$ and n is a real number, then;

$$log_{a}mn = nlog_{a}m$$

A straightforward way to explain Benford's Law is by examining the base-10 logarithms of numbers in a dataset. If the fractional parts of these logarithms are uniformly distributed within the interval $[0, 1]$ then the dataset is likely to follow Benford's Law.

To understand this, consider that a number xx begins with the digit d if and only if:

$$log_{10}d \leq log_{10}(x) < log_{10}(d+1)$$

This means the fractional part of $log_{10}x$ falls within an interval of length,

$$log_{10}(d+1) - log_{10}(d) = log_{10}\frac{d+1}{d}$$

Thus, the probability of a number starting with a digit d is directly related to this logarithmic interval, which is the basis of Benford's Law.

B.2 Scale invariance Deviation:

The following proof, originally presented by Daniel Cohen in 1976, builds on the concept of scale invariance introduced by Roger Pinkham in 1961, to derive Benford's Law as a discrete distribution.

Cohen's approach evaluates scale invariance by multiplying data sets by a factor of 2, showing the probabilities of the leading digits being 1. Then, he multiplies the

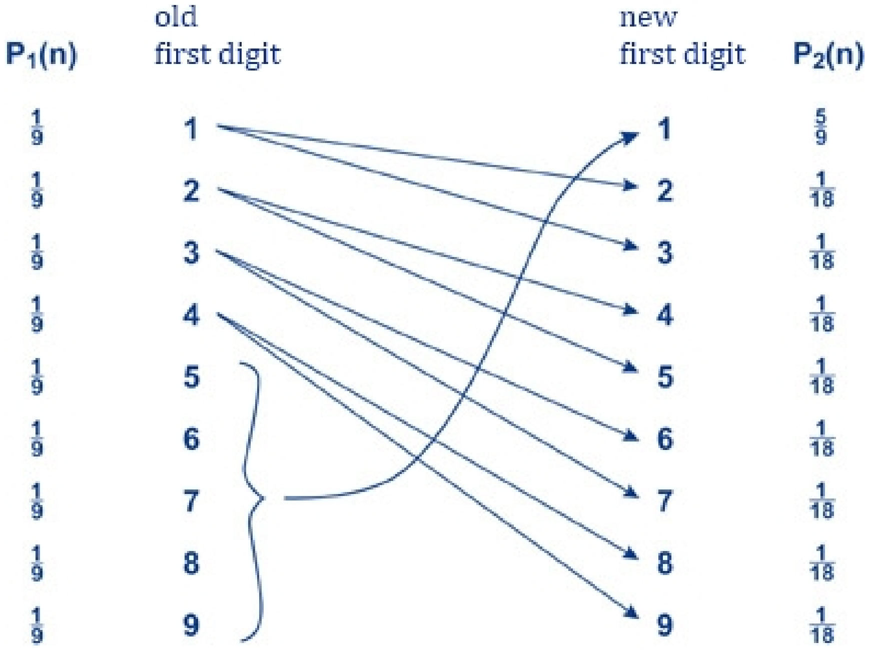

FIGURE B.1: Scale invariance for first digit test

data by a factor of 3 to demonstrate the probabilities of the leading digit being 2. The same method can be applied to derive the probabilities for the other leading digits.

The accompanying chart illustrates how numbers beginning with 5, 6, 7, 8, and 9, when doubled, often result in a number starting with 1. Similarly, numbers initially beginning with 1 typically transform into numbers starting with 2 or 3 upon doubling. As numbers with leading digits 2, 3, and 4 are doubled, their leading digits shift to 4 or 5, 6 or 7, and 8 or 9, respectively. The chart also highlights that the probabilities deviate from the uniform distribution expected under scale invariance. Instead of each digit occurring with equal frequency, there's a noticeable increase in the occurrence of the digit 1 after doubling, which contradicts the scale invariance expectation where $P_1 n = P_2 n$ should hold, where $P_1 n$ is the probability of the leading digit being n before doubling, and $P_2 n$ is the probability after doubling.

Instead of using the uniform probability of 1/9 for each digit, the more general probabilities $P(n), P(2), p(3)$, etc., are used. Due to the condition of scale invariance $P_1 n = P_2 n$, there's no distinction made between the probabilities of digits before and after doubling, so from now on, we use $P(n)$ to represent these probabilities. According to the chart, the relationship between specific probabilities can be generalized as $P(n) = p(2n) + p(2n+1)$. Additionally, when considering the first two digits together, /(P(n) represents the probability that these two digits match those of the two-digit number n.

Bibliography

[1] Benford, F., *The law of anomalous numbers. Proceedings of the American Philosophical Society*, **78(4)**, 551–572,1938

[2] Barabesi, L., Cerioli, A. Di Marzio,M.,*Statistical models and the Benford hypothesis: a unified framework*, **TEST 32**, 1479–1507 2023

[3] Berger A, Hill TP,*An Introduction to Benford's Law. Princeton, New Jersey: Princeton University Press*, **2015**

[4] Cheesing Lee, *Detecting numeric irregularities with Benford's Law*,May 15, 2015

[5] DAVID G. BANKS,*Moving Benford's Law from Art to Science*,2000

[6] Kossovsky, A.E.,*On the Mistaken Use of the Chi-Square Test in Benford's Law*,2021, 4, 419–453,*https://doi.org/10.3390/stats4020027*

[7] M.J. Nigrini, *Benford's Law: Applications for forensic accounting, auditing and fraud detection.*, **Chapter1**$_7$2012

[8] Morrow, J., *Benford's law, families of distributions and a test basis* **2010**

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

[10] Noether G. E. ,*Note on the Kolmogorov statistic in the discrete case Metrika* **7 (1)**,115–6,1963

[11] Steven J. Miller, *Benford's Law: Theory and Applications* **Chapter 1**,3-22,May 2015

[12] Raimi RA, *The first digit problem. The American Mathematical Monthly 83*, 1976,521–538

[13] R.; Lupi, C.,Some New Tests of Conformity with Benford's Law Stats 2021, 4, 745–761. https://doi.org/10.3390/stats 4030044

[14] R.j Sofytv USING BENFORD'S LAW TO DETECT FRAUD , *(NO. 02-5410)*