© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Chapter One

A Quick Introduction to Benford's Law

Steven J. Miller[1]

The history of Benford's Law is a fascinating and unexpected story of the interplay between theory and applications. From its beginnings in understanding the distribution of digits in tables of logarithms, the subject has grown enormously. Currently hundreds of papers are being written by accountants, computer scientists, engineers, mathematicians, statisticians and many others. In this chapter we start by stating Benford's Law of digit bias and describing its history. We discuss its origins and give numerous examples of data sets that follow this law, as well as some that do not. From these examples we extract several explanations as to the prevalence of Benford's Law, which are described in greater detail later in the book. We end by quickly summarizing many of the diverse situations in which Benford's Law holds, and why an observation that began in looking at the wear and tear in tables of logarithms has become a major tool in subjects as diverse as detecting tax fraud and building efficient computers. We then continue in the next chapters with rigorous derivations, and then launch into a survey of some of the many applications. In particular, in the next chapter we put Benford's Law on a solid foundation. There we explore several different categorizations of Benford's Law, and rigorously prove that certain systems satisfy these conditions.

1.1 OVERVIEW

We live in an age when we are constantly bombarded with massive amounts of data. Satellites orbiting the Earth daily transmit more information than is in the entire Library of Congress; researchers must quickly sort through these data sets to find the relevant pieces. It is thus not surprising that people are interested in patterns in data. One of the more interesting, and initially surprising, is Benford's Law on the distribution of the first or the leading digits.

In this chapter we concentrate on a mostly non-technical introduction to the subject, saving the details for later. Before we can describe the law, we must first set notation. At some point in secondary school, we are introduced to **scientific notation**: any positive number x may be written as $S(x) \cdot 10^k$, where $S(x) \in [1, 10)$ is the **significand** and k is an integer (called the **exponent**). The integer part of the

[1]Department of Mathematics and Statistics, Williams College, Williamstown, MA 01267. The author was partially supported by NSF grants DMS0970067 and DMS1265673.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

significand is called the **leading digit** or the **first digit**. Some people prefer to call $S(x)$ the mantissa and not the significand; unfortunately this can lead to confusion, as the **mantissa** is the fractional part of the logarithm, and this quantity too will be important in our investigations. As always, examples help clarify the notation. The number 1701.24601 would be written as $1.70124601 \cdot 10^3$ in scientific notation. The significand is 1.70124601, the exponent is 3 and the leading digit is 1. If we take the logarithm base 10, we find $\log_{10} 1701.24601 \approx 3.2307671196444460726$, so the mantissa is approximately .2307671196444460726.

There are many advantages to studying the first digits of a data set. One reason is that it helps us compare apples and apples and not apples and oranges. By this we mean the following: two different data sets could have very different scales; one could be masses of subatomic particles while another could be closing stock prices. While the units are different and the magnitudes differ greatly, every number has a unique leading digit, and thus we can compare the distribution of the first digits of the two data sets.

The most natural guess would be to assert that for a generic data set, all numbers are equally likely to be the leading digit. We would then posit that we should observe about 11% of the time a leading digit of 1, 2, . . . , 9 (note that we would guess each number occurs one-ninth of the time and not one-tenth of the time, as 0 is the leading digit for only one number, namely 0). The content of Benford's Law is that this is frequently not so; specifically, in many situations we expect the leading digit to be d with probability approximately $\log_{10}\left(\frac{d+1}{d}\right)$, which means the probability of a first digit of 1 is about 30% while a first digit of 9 happens about 4.6% of the time.

1.2 NEWCOMB

Though it is called Benford's Law, he was not the first to observe this digit bias. Our story begins with the astronomer–mathematician Simon Newcomb, who observed this behavior more than 50 years before Benford. Newcomb was born in Nova Scotia in 1835 and died in Washington, DC in 1909. In 1881 he published a short article in the American Journal of Mathematics, *Note on the Frequency of Use of the Different Digits in Natural Numbers* (see [New]). The article begins,

> That the ten digits do not occur with equal frequency must be evident to any one 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.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

> As natural numbers occur in nature, they are to be considered as the ratios of quantities. Therefore, instead of selecting a number at random, we must select two numbers, and inquire what is the probability that the first significant digit of their ratio is the digit n. To solve the problem we may form an indefinite number of such ratios, taken independently; and then must make the same inquiry respecting their quotients, and continue the process so as to find the limit towards which the probability approaches.

In this short article two very important properties of the distribution of digits are noted. The first is that all digits are not equally likely. The article ends with a quantification of how oftener the first digit is a 1 than a 9, with Newcomb stating,

> The law of probability of the occurrence of numbers is such that all mantissæ of their logarithms are equally probable.

Specifically, Newcomb gives a table (see Table 1.1) for the probabilities of first and second digits.

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 conjecture for the probabilities of observing a first digit of d or a second digit of d; all probabilities are reported to four decimal digits.

The second key observation of his paper is noting the importance of scale. The numerical value of a physical quantity clearly depends on the scale used, and thus Newcomb suggests that the correct items to study are ratios of measurements.

1.3 BENFORD

The next step forward in studying the distribution of the leading digits of numbers was Frank Benford's *The Law of Anomalous Numbers*, published in the Proceedings of the American Philosophical Society in 1938 (see [Ben]). In addition to advancing explanations as to why digits have this distribution, he also presents some justification as to why this is a problem worthy of study.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

> It has been observed that the pages of a much used table of common logarithms show evidences of a selective use of the natural numbers. The pages containing the logarithms of the low numbers 1 and 2 are apt to be more stained and frayed by use than those of the higher numbers 8 and 9. Of course, no one could be expected to be greatly interested in the condition of a table of logarithms, but the matter may be considered more worthy of study when we recall that the table is used in the building up of our scientific, engineering, and general factual literature. There may be, in the relative cleanliness of the pages of a logarithm table, data on how we think and how we react when dealing with things that can be described by means of numbers.

Benford studied the distribution of leading digits of 20 sets of data, including rivers, areas, populations, physical constants, mathematical sequences (such as $\sqrt{n}$, $n!$, n^2, ...), sports, an issue of Reader's Digest and the first 342 street addresses given in the (then) current American Men of Science. We reproduce his observations in Table 1.2.

Title	1	2	3	4	5	6	7	8	9	Count
Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
n^{-1}, $\sqrt{n}$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
Digest	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
n, n^2, ..., $n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average	30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Benford's Law	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6	

Table 1.2 Distribution of leading digits from the data sets of Benford's paper [Ben]; the amalgamation of all observations is denoted by "Average." Note that the agreement with Benford's Law is better for some examples than others, and the amalgamation of all examples is fairly close to Benford's Law.

Benford's paper contains many of the key observations in the subject. One of the most important is that while individual data sets may fail to satisfy Benford's Law, amalgamating many different sets of data leads to a new sequence whose behavior

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

is typically closer to Benford's Law. This is seen both in the row corresponding to n, $n^2, \ldots$ (where we can prove that each of these is non-Benford) as well as in the average over all data sets.

Benford's article suffered a much better fate than Newcomb's paper, possibly in part because it immediately preceded a physics article by Bethe, Rose and Smith on the multiple scattering of electrons. Whereas it was decades before there was another article building on Newcomb's work, the next article after Benford's paper was six years later (by S. A. Goutsmit and W. H. Furry, *Significant Figures of Numbers in Statistical Tables*, in Nature), and after that the papers started occurring more and more frequently. See Hurlimann's extensive bibliography [Hu] for a list of papers, books and reports on Benford's Law from 1881 to 2006, as well as the online bibliography maintained by Arno Berger and Ted Hill [BerH2].

1.4 STATEMENT OF BENFORD'S LAW

We are now ready to give precise statements of Benford's Law.

Definition 1.4.1 (Benford's Law for the Leading Digit). *A set of numbers satisfies Benford's Law for the Leading Digit if the probability of observing a first digit of d is $\log_{10}\left(\frac{d+1}{d}\right)$.*

While clean and easy to state, the above definition has several problems when we apply it to real data sets. The most glaring is that the numbers $\log_{10}\left(\frac{d+1}{d}\right)$ are irrational. If we have a data set with N observations, then the number of times the first digit is d must be an integer, and hence the observed frequencies are always rational numbers.

One solution to this issue is to consider only infinite sets. Unfortunately this is not possible in many cases of interest, as most real-world data sets are finite (i.e., there are only finitely many counties or finitely many trading days). Thus, while Definition 1.4.1 is fine for mathematical investigations of sequences and functions, it is not practical for many sets of interest. We therefore adjust the definition to

Definition 1.4.2 (Benford's Law for the Leading Digit (Working Definition)). *We say a data set satisfies Benford's Law for the Leading Digit if the probability of observing a first digit of d is approximately $\log_{10}\left(\frac{d+1}{d}\right)$.*

Note that the above definition is vague, as we need to clarify what is meant by "approximately." It is a non-trivial task to find good statistical tests for large data sets. The famous and popular chi-square tests, for example, frequently cannot be used with extensive data sets as this test becomes very sensitive to small deviations when there are many observations. For now, we shall use the above definition and interpret "approximately" to mean a good visual fit. This approach works quite well for many applications. For example, in Chapter 8 we shall see that many corporate and other financial data sets follow Benford's Law, and thus if the distribution is visually far from Benford, it is quite likely that the data's integrity has been compromised.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Finally, instead of studying just the leading digit we could study the entire significand. Thus in place of asking for the probability of a first digit of 1 or 2 or 3, we now ask for the probability of observing a significand between 1 and 2, or between π and e. This generalization is frequently called the **Strong Benford's Law**.

Definition 1.4.3 (Strong Benford's Law for the Leading Digits (Working Definition)). *We say a data set satisfies the Strong Benford's Law if the probability of observing a significand in* $[1, s)$ *is* $\log_{10} s$.

Note that Strong Benford behavior implies Benford behavior; the probability of a first digit of d is just the probability the significand is in $[d, d+1)$. Writing $[d, d+1)$ as $[1, d+1)\backslash[1, d)$, we see this probability is just $\log_{10}(d+1) - \log_{10} d = \log_{10} \frac{d+1}{d}$.

1.5 EXAMPLES AND EXPLANATIONS

In this section we briefly give some explanations for why so many different and diverse data sets satisfy Benford's Law, saving for later chapters more detailed explanation. It's worthwhile to take a few minutes to reflect on how Benford's Law was discovered, and to see whether or not similar behavior might be lurking in other systems. The story is that Newcomb was led to the law by observing that the pages in logarithm tables corresponding to numbers beginning with 1 were significantly more worn than the pages corresponding to numbers with higher first digit. A reasonable explanation for the additional wear and tear is that numbers with a low first digit are more common than those with a higher first digit. It is thus quite fortunate for the field that there were no calculators back then, as otherwise the law could easily have been missed. Though few (if any) of us still use logarithm tables, it is possible to see a similar phenomenon in the real world today. Our analysis of this leads to one of the most important theorems in probability and statistics, the Central Limit Theorem, which plays a role in understanding the ubiquity of Benford's Law.

Instead of looking at logarithm tables, we can look at the steps in an old building, or how worn the grass is on college campuses. Assuming the steps haven't been replaced and that there is a reasonable amount of traffic in and out of the building, then lots of people will walk up and down these stairs. Each person causes a small amount of wear and tear on the steps; though each person's contribution is small, if there are enough people over a long enough time period then the cumulative effect will be visually apparent. Typically the steps are significantly more worn towards the center and less so as one moves towards the edges. A little thought suggests the obvious answer: people typically walk up the middle of a flight of stairs unless someone else is coming down. Similar to carbon dating, one could attempt to determine the age of a building by the indentation of the steps. Looking at these patterns, we would probably see something akin to the normal distribution, and if we were fortunate we might "discover" the Central Limit Theorem. There are many other examples from everyday life. We can also observe this in looking at lawns. Everyone knows the shortest distance between two points is a line, and people frequently leave the sidewalks and paths and cut across the grass, wearing

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

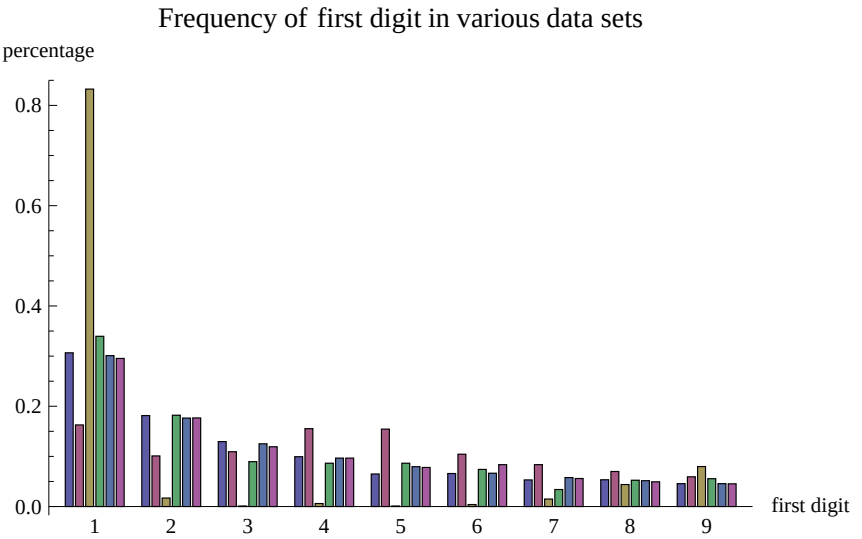

Figure 1.1 Frequencies of leading digits for (a) U.S. county populations (from 2000 census); (b) U.S. county land areas in square miles (from 2000 census); (c) daily volume of NYSE trades from 2000 through 2003; (d) fundamental constants (from NIST); (e) first 3219 Fibonacci numbers; (f) first 3219 factorials. Note the census data includes Puerto Rico and the District of Columbia.

it down to dirt in some places and leaving it untouched in others. Another example is to look at keyboards, and compare the well-worn "E" to the almost pristine "Q." Or the wear and tear on doors. The list is virtually endless.

In Figure 1.1 we look at the leading digits of the several "natural" data sets. Four arise from the real world, coming from the 2000 census in the United States (population and area in square miles of U.S. counties), daily volumes of transactions on the New York Stock Exchange (NYSE) from 2000 through 2003 and the physical constants posted on the homepage of the National Institute for Standards and Technology (NIST); the remaining two data sets are popular mathematical sequences: the first 3219 Fibonacci numbers and factorials (we chose this number so that we would have as many entries as we do counties).

If these are "generic" data sets, then we see that no one law describes the behavior of each set. Some of the sets are quite close to following Benford's Law, others are far off; none are close to having each digit equally likely to be the leading digit. Except for the second and third sets, the rest of the data behaves similarly; this is easier to see if we remove these two examples, which we do in Figure 1.2.

Before launching into explanations of why so many data sets are Benford (or at least close to it), it's worth briefly remarking why many are not. There are several reasons and ways a data set can fail to be Benford; we quickly introduce some of these reasons now, and expand on them more when we advance explanations for Benford's Law below. For example, imagine we are recording hourly temperatures

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

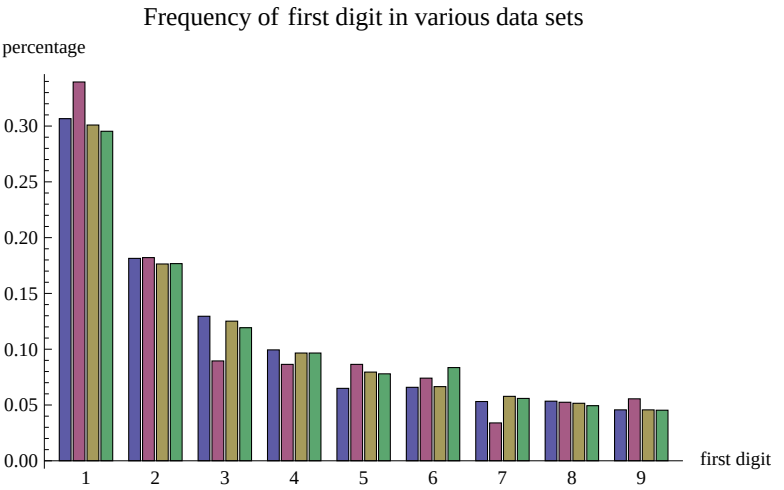

Figure 1.2 Frequencies of leading digits for (a) U.S. county populations (from 2000 census); (b) fundamental constants (from NIST); (c) first 3219 Fibonacci numbers; (d) first 3219 factorials. Note the census data includes Puerto Rico and the District of Columbia.

in May at London Heathrow Airport. In Fahrenheit the temperatures range from lows of around 40 degrees to highs of around 80. As all digits are not accessible, it's impossible to be Benford, though perhaps *given this restriction, the relative probabilities of the digits are Benford*.

For another issue, we have many phenomena that are given by specific, concentrated distributions that will not be Benford. The Central Limit Theorem is often a good approximation for the behavior of numerous processes, ranging from heights and weights of people to batting averages to scores on exams. In these situations we clearly do not expect Benford behavior, though we will see below that processes whose *logarithms* are normally distributed (with large standard deviations) are close to Benford.

Thus, in looking for data sets that are close to Benford, it is natural to concentrate on situations where the values are not given by a distribution concentrated in a small interval. We now explore some possibilities below.

1.5.1 The Spread Explanation

We drew the examples in Figure 1.1 from very different fields; why do so many of them behave similarly, and why do others violently differ? While the first question still confounds researchers, we can easily explain why two data sets had such different behavior, and this reason has been advanced by many as a source of Benford's Law (though there are issues with it, which we'll comment on shortly). Let's look at the first two sets of data: the population in U.S. counties in 2000 and daily volume of the NYSE from 2000 through 2003. You can see from the histogram in

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

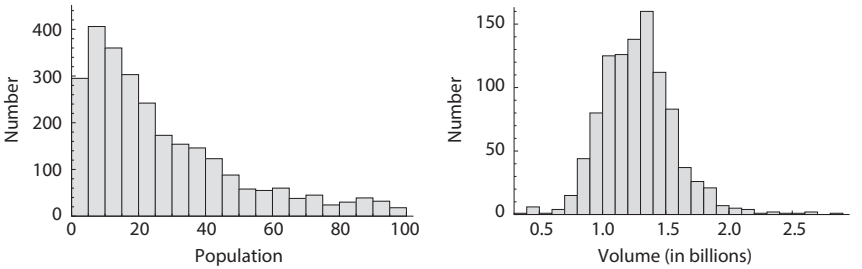

Figure 1.3 (Left) The population (in thousands) of U.S. counties under 250,000 (which is about 84% of all counties). (Right) The daily volume of the NYSE from 2000 through 2003. Note the population spans two orders of magnitude while the stock volumes are mostly within a factor of 2 of each other.

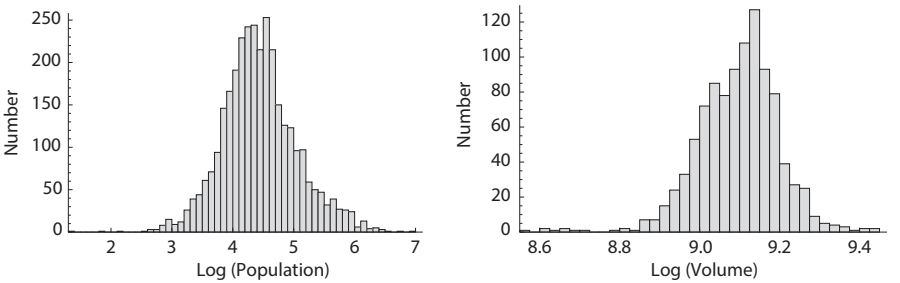

Figure 1.4 (Left) The population of U.S. counties. (Right) The daily volume of the NYSE from 2000 through 2003.

Figure 1.3 the stock market transactions are clustered around one value and span only one order of magnitude. Thus it is not surprising that there is little variation in these first digits. For the county populations, however, the data is far more spread out. These effects are clearer if we look at a histogram of the log-plot of the data, which we do in Figure 1.4. A detailed analysis of the other data sets shows similar behavior; the four data sets that behave similarly are spread out on a logarithmic plot over several orders of magnitude, while the two sets that exhibit different behavior are more clustered on a log-plot.

Our discussion above leads to our first explanation for Benford's Law, the **spread hypothesis**. The spread hypothesis states that if a data set is distributed over several orders of magnitude, then the leading digits will approximately follow Benford's Law. Of course, a little thought shows that we need to assume far more than the data just being spread out over several orders of magnitude. For example, if our set of observations were

$$\{1,\ 10,\ 100,\ 1000,\ \ldots,\ 10^{2015}\}$$

then clearly it is non-Benford, even though it does cover over 2000 orders of magnitude! As remarked above, our purpose in this introduction is to just briefly intro-

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

duce the various ideas and approaches, saving the details for later. There are many issues with the **spread hypothesis**; see Chapter 2 and [BerH3] for an excellent analysis of these problems.

1.5.2 The Geometric Explanation

Our next attempt to explain the prevalence of Benford's Law goes back to Benford's paper [Ben], whose second part is titled *Geometric Basis of the Law*. The idea is that if we have a process with a constant growth rate, then more time will be spent at lower digits than higher digits. For definiteness, imagine we have a stock that increases at 4% per year. The amount of time it takes to move from \$1 to \$2 is the same as it would take to move from \$10,000 to \$20,000 or from \$100,000,000 to \$200,000,000. If n_d is the number of years it takes to move from d dollars to $d+1$ dollars then $d \cdot (1.04)^{n_d} = (d+1)$, or

$$n_d \;=\; \frac{\log\left(\frac{d+1}{d}\right)}{\log 1.04}. \tag{1.1}$$

In Table 1.3 we consider the (happy) situation of a stock that rises 4% each and every year. Notice that it takes over 17 years to move from being worth \$1 to being worth \$2, but less than 3 years to move from being worth \$9 to \$10.

First digit	Years	Percentage of time	Benford's Law
1	17.6730	0.30103	0.30103
2	10.3380	0.17609	0.17609
3	7.3350	0.12494	0.12494
4	5.6894	0.09691	0.09691
5	4.6486	0.07918	0.07918
6	3.9303	0.06695	0.06695
7	3.4046	0.05799	0.05799
8	3.0031	0.05115	0.05115
9	2.6863	0.04576	0.04576

Table 1.3 How long the first digit of a stock has leading digit d, given that the stock rises 4% each year. It takes the stock approximately 58.7084 years to increase from \$1 to \$10.

A little algebra shows that this implies Benford behavior. If n is the amount of time it takes to move from \$1 to \$10, then $1 \cdot (1.04)^n = 10$ or $n = \frac{\log 10}{\log 1.04}$. Thus by (1.1), we see the percentage of the time spent with a first digit of d is

$$\frac{\log\left(\frac{d+1}{d}\right)}{\log 1.04} \Bigg/ \frac{\log 10}{\log 1.04} \;=\; \frac{\log\left(\frac{d+1}{d}\right)}{\log 10} \;=\; \log_{10}\left(\frac{d+1}{d}\right), \tag{1.2}$$

which is just Benford's Law! There is nothing special about 4%; the same analysis works in general *provided* that at each moment we grow by the same, fixed rate. The

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

analysis is more interesting if at each instance the growth percentage is a random variable, say drawn from a Gaussian. For more on such processes see Chapter 6.

This is not an isolated example. Many natural and mathematical phenomena are governed by geometric growth. Examples range from radioactive decay and bacteria populations to the Fibonacci numbers. One reason for this is that solutions to many difference equations are given by linear combinations of geometric series; as difference equations are just discrete analogues of differential equations, it is thus not surprising that they model many situations. For example, the Fibonacci numbers satisfy the second order linear recurrence relation

$$F_{n+2} \;=\; F_{n+1} + F_n. \tag{1.3}$$

Once the first two Fibonacci numbers are known, the recurrence (1.3) determines the rest. If we start with $F_0 = 0$ and $F_1 = 1$, we find $F_2 = 1$, $F_3 = 2$, $F_4 = 3$, $F_5 = 5$ and so on. Moreover, there is an explicit formula for the nth term, namely

$$F_n \;=\; \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n; \tag{1.4}$$

known as Binet's formula; generalizations of it hold for solutions to linear recurrence relations. As $\left|\frac{1+\sqrt{5}}{2}\right| > 1$ and $\left|\frac{1-\sqrt{5}}{2}\right| < 1$, for large n this implies $F_n \approx \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n$. Note that $F_{n+1} \approx \frac{1+\sqrt{5}}{2}F_n$, or $F_{n+1} \approx 1.61803F_n$. This means that the Fibonacci numbers are well approximated by what would be a highly desirable stock rising about 61.803% each year, and hence by our previous analysis it is reasonable to expect the Fibonacci numbers will be Benford as well.

While the discreteness of the Fibonacci numbers makes the analysis a bit more complicated than the continuous growth rate problem, a generalization of these methods proves that the Fibonacci numbers, as well as the solution to many difference equations, are Benford. Again, our purpose here is to merely provide some evidence as to why so many different, diverse systems satisfy Benford's Law. It is not the case that every recurrence relation leads to Benford behavior. To see this, consider $a_{n+2} = 2a_{n+1} - a_n$ with either $a_0 = a_1 = 1$ (which implies $a_n = 1$ for all n) or $a_0 = 0$ and $a_1 = 1$ (which implies $a_n = n$ for all n). While there are examples of recurrence relations that are non-Benford, a "generic" one will satisfy Benford's Law, and thus studying these systems provides another path to Benford.

1.5.3 The Scale-Invariance Explanation

For our next explanation, we return to a comment from Newcomb's [New] paper:

> As natural numbers occur in nature, they are to be considered as the ratios of quantities. Therefore, instead of selecting a number at random, we must select two numbers, and inquire what is the probability that the first significant digit of their ratio is the digit n.

The import of this comment is that the behavior should be independent of the units used. For example, if we look at the value of stocks in our portfolio then the magnitudes will change if we measure their worth in dollars or euros or yen or bars

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

of gold pressed latinum, though the physical quantities are unchanged. Similarly we can use the metric system or the (British) imperial system in measuring physical constants. As the universe doesn't care what units we use for our experiments, it is natural to expect that the distribution of leading digits should be unchanged if we change our units.

For definiteness, let's consider the areas of the countries in the world. There are almost 200 countries; if we measure area in square kilometers then about 28.49% have a first digit of 1 and 18.99% have a first digit of 2, while if we measure in square miles it is 34.08% have a first digit of 1 and 16.20% have a first digit of 2, which should be compared to the Benford probabilities of approximately 30.10% and 17.61%; one observes a similar closeness with the other digits.

The assumption that there *is* a distribution of the first digit and that this distribution is independent of scale implies the first digits follow Benford's Law. The analysis of this involves introducing a σ-algebra and studying scale-invariant probability measures on this space. Without going into these details now, we can at least show that Benford's Law is consistent with scale invariance.

Let's assume our data set satisfies the Strong Benford Law (see Definition 1.4.3). Then the probability the significand is in $[a, b] \subset [1, 10)$ is $\log_{10}(b/a)$. Assume now we rescale every number in our set by multiplying by a fixed constant C. For definiteness we take $C = \sqrt{3}$ and compute the probability that numbers in the scaled data set have leading digit 1. Note that multiplying $[1, 10)$ by $\sqrt{3}$ gives us the interval $[\sqrt{3}, 10\sqrt{3}) \approx [1.73, 17.32)$. The parts of this new interval with a leading digit of 1 are $[\sqrt{3}, 2)$ and $[10, 10\sqrt{3})$, which come from $[1, 2/\sqrt{3})$ and $[10/\sqrt{3}, 10)$. As we are assuming the strong form of Benford's Law, the probabilities of these two intervals are $\log_{10} \frac{2/\sqrt{3}}{1}$ and $\log_{10} \frac{10}{10\sqrt{3}}$. Summing yields the probability of the first digit of the scaled set being 1 is

$$\log_{10}\left(\frac{2/\sqrt{3}}{1}\right) + \log_{10}\left(\frac{10}{10\sqrt{3}}\right) \;=\; \log_{10} 2,$$

which is the Benford probability! A similar analysis works for the other leading digits and other choices of C.

We close this section by noting that scale invariance fits naturally with the other explanations introduced to date. If our initial data set were spread out over several orders of magnitude, so too would the scaled data. Similarly, if we return to our hypothetical stock increasing by 4% per year, the effect of changing the units of our currency can be viewed as changing our principal; however, what governs how long our stock spends with a leading digit of d is not the principal but rather the rate of growth, and that is unchanged.

1.5.4 The Central Limit Explanation

We need to introduce some machinery for our last heuristic explanation. If $y \geq 0$ is a real number, by $y \bmod 1$ we mean the fractional part of y. Other notations for this are $\{y\}$ or $y - \lfloor y \rfloor$. If $y < 0$ then $y \bmod 1$ is $1 - (-y \bmod 1)$. In other words, $y \bmod 1$ is the unique number in $[0, 1)$ such that $y - (y \bmod 1)$ is an integer. Thus $3.14 \bmod 1$ is .14, while $-3.14 \bmod 1$ is .86. We say y **modulo** 1 for $y \bmod 1$.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Recall that any positive number x may be written in **scientific notation** as $x = S(x) \cdot 10^k$, where $S(x) \in [1, 10)$ and k is an integer. The real number $S(x)$, called the **significand**, encodes all the information about the digits of x; the effect of k is to specify the decimal point's location. Thus, if we are interested in either the first digit or the significand, the value of k is immaterial. This suggests that rather than studying our data as given, it might be worthwhile to transform the data as follows:

$$x \ \mapsto \ \log_{10} x \bmod 1. \tag{1.5}$$

A little algebra shows that two positive numbers have the same leading digits if and only if their signficands have the same first digit. Thus if we have a set of values $\{x_1, x_2, x_3, \dots\}$ then the subset with leading digit d is $\{x_i : S(x_i) \in [d, d+1)\}$, which is equivalent to $\{x_i : \log_{10} S(x_i) \in [\log_{10} d, \log_{10}(d+1))\}$.

This innocent-looking reformulation turns out to be not only one of the most fruitful ways of exploring Benford's Law, but also highlights what is going on. We first explain the new perspective gained by transforming the data. According to Benford's Law, the probability of observing a first digit of d is $\log_{10} \frac{d+1}{d}$. This is $\log_{10}(d+1) - \log_{10} d$, which is the length of the interval $[\log_{10} d, \log_{10}(d+1))$! In other words, consider a data set satisfying Benford's Law, and transform the set as in (1.5). The new set lives in $[0, 1)$ and is uniformly distributed there. Specifically, the probability that we have a value in the interval $[\log_{10} d, \log_{10}(d+1))$ is the length of that interval.

While it may not seem natural to take the logarithm base 10 of each number, and then look at the result modulo 1, under such a process the resulting values are uniformly distributed if the initial set obeys Benford's Law. Another way of looking at this is that there is a natural transformation which takes a set satisfying Benford's Law and returns a new set of numbers that is uniformly distributed.

We briefly comment on why this is a natural process. We replace x with $\log_{10} x \bmod 1$. If we write $x = S(x) \cdot 10^k$, then $\log_{10} x \bmod 1$ is just $\log_{10} S(x)$. Thus taking the logarithm modulo 1 is a way to get our hands on the significand (actually, its logarithm), which is what we want to understand. While the logarithm function is a nice function, removing the integer part *in general* is messy and leads to complications; however, there is a very important situation where it is painless to remove the integer part. Recall the exponential function

$$e(x) \ := \ e^{2\pi i x} \ = \ \cos(2\pi x) + i \sin(2\pi x), \tag{1.6}$$

where $i = \sqrt{-1}$. As $e(x+1) = e(x)$, we see

$$e(x \bmod 1) \ = \ e(x). \tag{1.7}$$

The utility of the above becomes apparent when we apply Fourier analysis. In Fourier analysis one uses sines, cosines or exponential functions to understand more complicated functions. From our analysis above, we may either include the modulo 1 or not in the argument of the exponential function. While we will elaborate on this at great length later, the key takeaway is that the transformed data is ideally suited for Fourier analysis.

We can now sketch how this is related to Benford's Law. There are many data sets in the world whose values are the product of numerous measurements. For

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

example, the monetary value of a gold brick is a product of the brick's length, width, height, density and value of gold per pound. Imagine we have some quantity X which is a product of n values, so

$$X \;=\; X_1 \cdot X_2 \cdot \ldots \cdot X_n.$$

We assume the X_i's are nice random variables. From our discussion above, to show that X obeys Benford's Law it suffices to know that the distribution of the logarithm of X modulo 1 is uniformly distributed. Thus we are led to study

$$\log_{10} X \;=\; \log_{10}(X_1 \cdot X_2 \cdot \ldots \cdot X_n) \;=\; \log_{10} X_1 + \cdots + \log_{10} X_n.$$

By the Central Limit Theorem, if n is large then the above sum is approximately normally distributed, and the variance will grow with n; however, what we are really interested in is not this sum but rather this sum modulo 1:

$$\log_{10} X \bmod 1 \;=\; (\log_{10} X_1 + \cdots + \log_{10} X_n) \bmod 1.$$

A nice computation shows that as the variance σ tends to infinity, if we look at the probability density of a normal with variance σ modulo 1 then that is approximately uniformly distributed on $[0, 1]$. Explicitly, let Y be normally distributed with some mean μ and very large variance σ. If we look at the probability density of the new random variable $Y \bmod 1$, then this is approximately uniformly distributed on $[0, 1)$. This means that the probability that $Y \in [\log_{10} d, \log_{10}(d+1))$ is just $\log_{10}(d+1) - \log_{10} d$, or $\log_{10} \frac{d+1}{d}$; however, note that these are just the Benford probabilities!

While we have chosen to give the argument for multiplying random variables, similar results hold for other combinations (such as addition, exponentiation, etc.). The Central Limit Theorem is lurking in the background, and if we adjust our viewpoint we can see its effect.

1.6 QUESTIONS

Our goal in this book is to explain the prevalence of Benford's Law, and discuss its implications and applications. The question of leading digits is but one of many that we could ask. There are many generalizations; below we state the two most common.

1. *Instead of studying the distribution of the first digit, we may study the distribution of the first two, three, or more generally the significand, of our number. The Strong Benford's Law is that the probability of observing a significand of at most s is $\log_{10} s$.*

2. *Instead of working in base 10, we may work in base B, in which case the Benford probabilities become $\log_B \left(\frac{d+1}{d}\right)$ for the distribution of the first digit, and $\log_B s$ for a significand of at most s.*

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

Incorporating these two generalizations, we are led to our final definition of Benford's Law.

Definition 1.6.1 (Strong Benford's Law Base B). *A data set satisfies the Strong Benford's Law Base B if the probability of observing a significand of at most s in base B is $\log_B s$. We shall often refer to the distribution of just the first digit as Benford's Law, as well as the distribution of the entire significand.*

We end the introduction by briefly summarizing the goals of this book and what follows. We address two central questions:

1. *Which data sets (mathematical expressions, physical data, financial transactions) follow this law, and why?*

2. *What are the practical implications of this law?*

There are several different arguments for the first question, depending on the structure of the data. Our studies will show that the answer is deeply connected to results in subjects ranging from probability to Fourier analysis to dynamical systems to number theory. We shall develop enough of these topics for our investigations, recalling standard results in each when needed.

The second question leads to many surprising characters entering the scene. The reason Benford's Law is not just a curiosity of pure mathematics is due to the wealth of applications, in particular to data integrity and fraud tests. There have (sadly) been numerous examples of researchers and corporations tinkering with data; if undetected, the consequences could be severe, ranging from companies not paying their fair share of taxes, to unsafe medical treatments being approved, to unscrupulous researchers being funded at the expense of their honest peers, to electoral fraud and the effective disenfranchisement of voters. With a large enough data set, the laws of probability and statistics state that certain patterns should emerge. Some of these consequences are well known, and thus are easily incorporated by people modifying data. For example, while everyone knows that if you simulate flipping a fair coin 1,000,000 times then there should be about 500,000 heads, fewer know how likely it is to have 100 consecutive heads in the sequence of tosses. The situation is similar with Benford's Law. Almost anyone unfamiliar with Benford's Law would, if asked to simulate data, create a set where either the first digits are equally likely to be anything from 1 to 9, or else clustered around 5. As many real-world data sets follow Benford's Law, this leads to a quick and easy test for fraud. Such tests are now routinely used by the IRS to detect tax fraud, while generalizations may be used in the future to detect whether or not an image has been modified.

For general queries, contact webmaster@press.princeton.edu

© Copyright, Princeton University Press. No part of this book may be distributed, posted, or reproduced in any form by digital or mechanical means without prior written permission of the publisher.

What better way to end the introduction than with notes from a talk that Frank Benford gave on the law that now bears his name! While this was one of the earliest talks in the subject, it was by no means the last. As the online bibliography [BerH2] shows, Benford's Law has become a very active research area with numerous applications across disciplines, many of which are described in the following chapters. Enjoy!

For general queries, contact webmaster@press.princeton.edu