Theory and Applications of Benford's Law

Theory and Applications of Benford's Law

Steven J. Miller (senior editor)
Arno Berger and Ted Hill
(editors)

PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

Published by Princeton University Press, 41 William Street, Princeton, New Jersey 08540

In the United Kingdom: Princeton University Press, 3 Market Place, Woodstock, Oxfordshire OX20 1SY

Library of Congress Cataloging-in-Publication Data

Miller, Steven J.
Theory and Applications of Benford's Law / Steven J. Miller et al.
p. cm.
Includes bibliographical references and index.
ISBN-13: TBD (cloth : alk. paper)
ISBN-10: TBD (cloth : alk. paper)
1. Probability and Statistics. II. Title.

TBD
TBD

British Library Cataloging-in-Publication Data is available

This book has been composed in Times Roman in LaTeX

The publisher would like to acknowledge the authors of this volume for providing the camera-ready copy from which this book was printed.

Printed on acid-free paper. ∞

pup.princeton.edu

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

DEDICATION FROM PREVIOUS BOOK: To our wives, for their patience and constant encouragement, and to our students and colleagues, whose insights and exuberance made this project both possible and enjoyable.
NEW: many people, how about also Frank Benford and Simon Newcomb for providing such an interesting and rewarding subject to study?

Contents

Foreword

TBD name
TBD address
TBD Date

Preface

VERY ROUGH PREFACE: Benford's law of digit bias appears in a variety of different disciplines, including probability and statistics, accounting, engineering, number theory, dynamical systems and computer science, to name just a few. Our purposes are to show students and researchers useful techniques from a variety of subjects, highlight the connections between the different areas, and encourage research and cross-departmental collaboration on these problems. To do this, we develop much of the general theory in the first few chapters (concentrating on the methods which are applicable to a variety of problems), and then conclude with numerous chapters on applications written by a world-expert in that field. Though there are common themes and methods throughout the applications, these chapters are self-contained, needing only the introductory chapters (and some standard material reviewed in the appendices). When appropriate, especially in the introductory chapters, we include exercises and open problems so that this book may be used as a textbook at either the undergraduate or graduate level. In particular, there are chapters on dynamical systems and Fourier analysis, so this book may be used as an introduction to these subjects with applications.

We are extremely grateful to Princeton University Press, especially to our editor Vickie Kearn and to **ADD OTHERS**, our production editor **TBD** and our copyeditor **TBD**, for all their help and aid, and to **ADD THANKS TO THOSE WHO FUNDED**.

The senior editor was partially supported by NSF Grant DMS0600848. Much of this book is based on a conference organized by Chaouki T. Abdallah, Gregory L. Heileman, Steven J. Miller and Fernando Pérez-González: Conference on the Theory and Applications of Benford's Law (December 16th - 18th, 2007, Santa Fe, NM). This conference was supported in part by Brown University, IEEE, NSF Grant DMS-0753043, New Mexico Consortium's Institute for Advanced Study, Universidade de Vigo and the University of New Mexico, and it is a pleasure to thank them and all the participants.

Steven J. Miller
Williamstown, MA
TBD month 2010

OTHER EDITORS?
LOCATION TBD

TBD month 2010

Notation

$\mathbb{W}$: the set of whole numbers: $\{1, 2, 3, 4, \dots\}$.

$\mathbb{N}$: the set of natural numbers: $\{0, 1, 2, 3, \dots\}$.

$\mathbb{Z}$: the set of integers: $\{\dots, -2, -1, 0, 1, 2, \dots\}$.

$\mathbb{Q}$: the set of rational numbers: $\{x : x = \frac{p}{q}, p, q \in \mathbb{Z}, q \neq 0\}$.

$\mathbb{R}$: the set of real numbers.

$\mathbb{C}$: the set of complex numbers: $\{z : z = x + iy,\ x, y \in \mathbb{R}\}$.

$\Re z$, $\Im z$: the real and imaginary parts of $z \in \mathbb{C}$; if $z = x+iy$, $\Re z = x$ and $\Im z = y$.

$x \equiv y \mod n$: there exists an integer a such that $x = y + an$.

$\forall$: for all.

$\exists$: there exists.

Big-Oh notation : $A(x) = O(B(x))$, read "$A(x)$ is of order (or big-Oh) $B(x)$", means there exists a $C > 0$ and an x_0 such that for all $x \geq x_0$, $|A(x)| \leq CB(x)$. This is also written $A(x) \ll B(x)$ or $B(x) \gg A(x)$.

Little-Oh notation : $A(x) = o(B(x))$, read "$A(x)$ is little-Oh of $B(x)$", means $\lim_{x\to\infty} A(x)/B(x) = 0$.

$|S|$ or $\#S$: number of elements in the set S.

$[x]$ or $\lfloor x \rfloor$: the greatest integer less than or equal to x, read "the floor of x".

$\{x\}$: the **fractional part** of x; note $x = [x] + \{x\}$.

supremum : given a sequence $\{x_n\}_{n=1}^{\infty}$, the supremum of the set, denoted $\sup_n x_n$, is the smallest number c (if one exists) such that $x_n \leq c$ for all n, and for any $\epsilon > 0$ there is some n_0 such that $x_{n_0} > c - \epsilon$. If the sequence has finitely many terms, the supremum is the same as the maximum value.

infimum : notation as above, the infimum of a set, denoted $\inf_n x_n$, is the largest number c (if one exists) such that $x_n \geq c$ for all n, and for any $\epsilon > 0$ there is some n_0 such that $x_{n_0} < c + \epsilon$. If the sequence has finitely many terms, the infimum is the same as the minimum value.

$\Box$: indicates the end of a proof.

PART 1
Introduction and Background Material (**POOR TITLE**)

Chapter One

Introduction

1.1 OVERVIEW

We live in an age where 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. The purpose of this book is to describe this law, discuss what sorts of data sets do and do not follow it, provide a theoretical explanation for this behavior, and finally discuss the multitude of applications this innocent seeming law has.

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 fix 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 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 $1.70124601 \cdot 10^3$. 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. If instead we chose .00729735257 we would find the significand is 7.29735257, the exponent is -3, the leading digit is 7 and $\log_{10} .00729735257 \approx$ -2.136834670397124824 so the mantissa is approximately .136834670397124824.

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 first digits of the two data sets.

This leads us to the central question of this book:

How do we expect the leading digits of a data set of positive

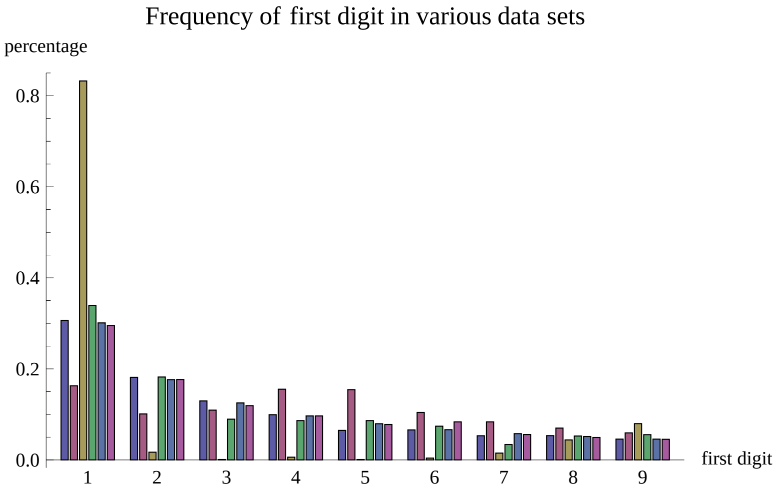


Figure 1.1 Frequencies of leading digits for (a) U.S. county populations (from 2000 census); (b) U.S. county land areas in miles^2 (from 2000 census); (c) daily volume of NYSE trades from 2000 through 2003; (d) fundamental constants (from NIST); (e) first 3219 Fibonacci numbers; (f) first 3219 factorials. Note the census data includes Puerto Rico and the District of Columbia.

numbers distributed?

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 of 1, 2, . . . , or 9 (note that we would guess each number occurs one-ninth of the time and not one-tenth of the time, as 0 is not allowed to be a leading digit).

Let's test this hypothesis by looking at some examples. 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 miles^2 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 our conjecture that each number occurs 11% of the time as the first digit is clearly false. We see a strong bias towards lower first digits. 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.

We drew these examples 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 sets had such different be-

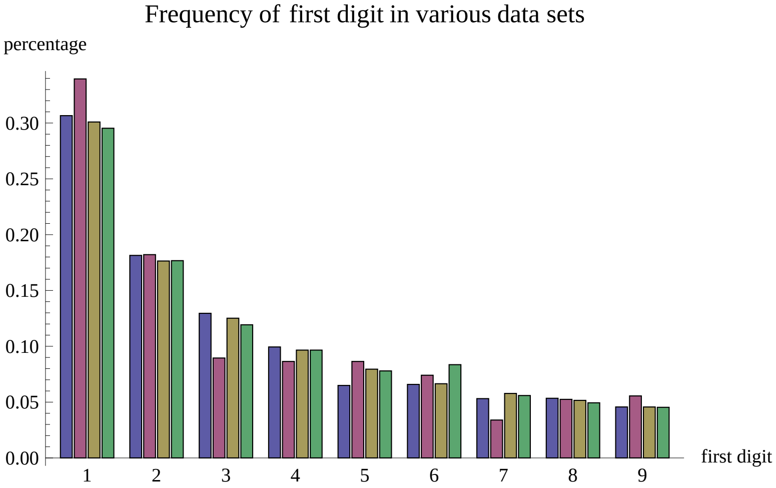


Figure 1.2 Frequencies of leading digits for (a) U.S. county populations (from 2000 census); (b) fundamental constants (from NIST); (c) first 3219 Fibonacci numbers; (d) first 3219 factorials. Note the census data includes Puerto Rico and the District of Columbia.

havior. 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 2004.

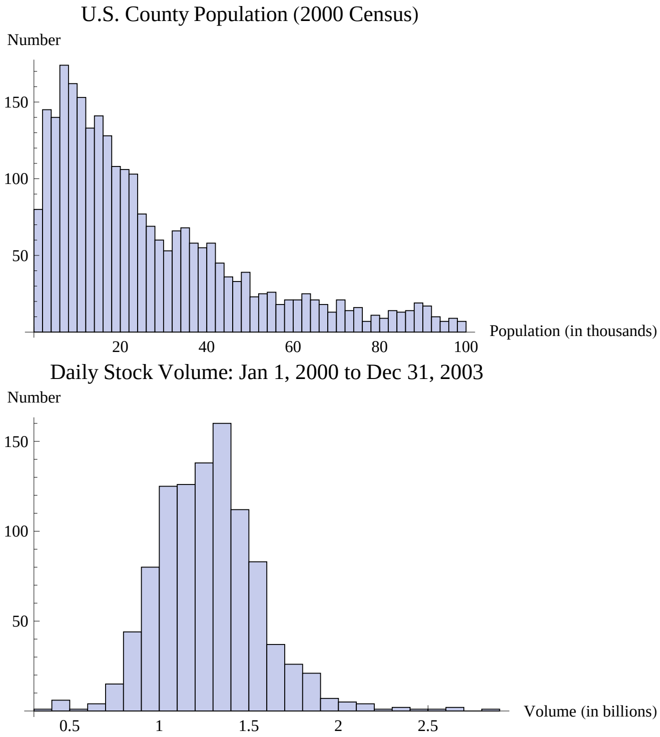
We note that the stock market transactions are clustered around one value, and thus not surprisingly there is little variation in the first digits. For the county populations, however, the data is far more spread out. These effects are clearer if we look at the log-plot of Figure 1.4.

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.

These and other examples (which will be described in greater detail in §1.2) led researchers to posit a law for the distribution of leading digits of many data sets. Though discovered by Newcomb in the 1880s, it is known as Benford's law as Benford's paper (from the 1930s) was responsible for disseminating these observations to a wide audience. We are led to the following definition.

Definition 1.1.1 (**Benford's Law for the Leading Digit (Theoretical Definition)**). *Benford's Law for the Leading Digit states that for many natural sets of data, the probability of observing a first digit of* d *is* $\log_{10}\left(\frac{d+1}{d}\right)$.

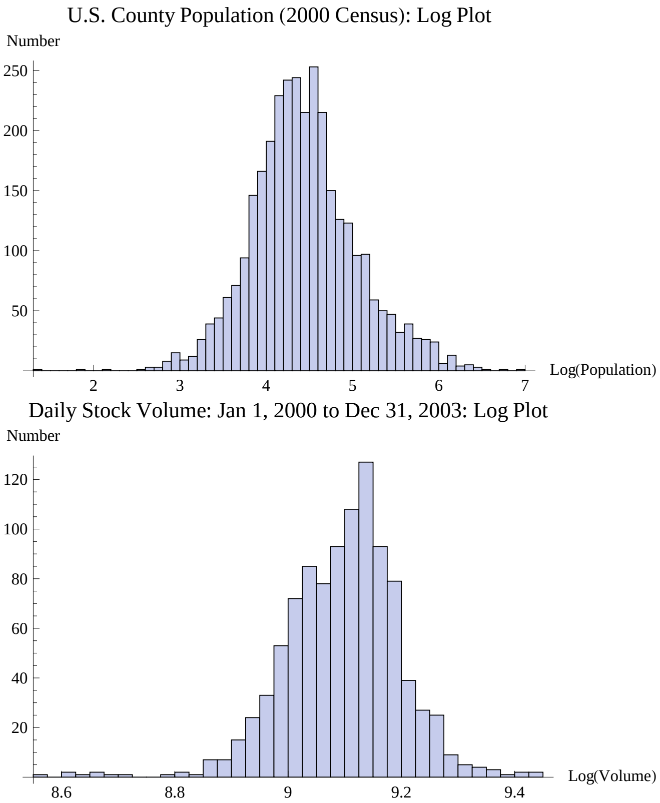
While clean and easy to state, the above definition has several problems when we apply it to real data sets. The most glaring is that the numbers $\log_{10}\left(\frac{d+1}{d}\right)$ are irrational. If we have a data set with N observations, then the number of times the first digit is d must be an integer,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



U.S. County Population (2000 Census)

Daily Stock Volume: Jan 1, 2000 to Dec 31, 2003

Figure 1.3 (Top) The population of U.S. counties under 250,000 (which is about 84% of all counties). (Bottom) The daily volume of the NYSE from 2000 through 2003.



U.S. County Population (2000 Census): Log Plot

Daily Stock Volume: Jan 1, 2000 to Dec 31, 2003: Log Plot

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

One solution to this issue is to only consider 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 1.1.1 is fine for mathematical investigations, or infinite mathematical sets (such as the set of all Fibonacci numbers or the values of $n!$), it is not practical for many sets of interest. We therefore adjust the definition to

Definition 1.1.2 (**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 **ADD CHAPTER REF** we shall see that many corporate and other financial data 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.

Armed with these definitions, we now compare the probabilities from the data sets of Figure 1.2 to the Benford probabilities in Figure 1.5. To aid the reader, we offset the Benford probabilities from the four data sets with a blank column. The visual fit is quite appealing.

As stated above, our goal in this book is to explain how universal this behavior is, and discuss its implications. 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. Benford's law becomes the probability of observing a significand of at most s is $\log_{10} s$.

2. Instead of working base 10, we may work 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.

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

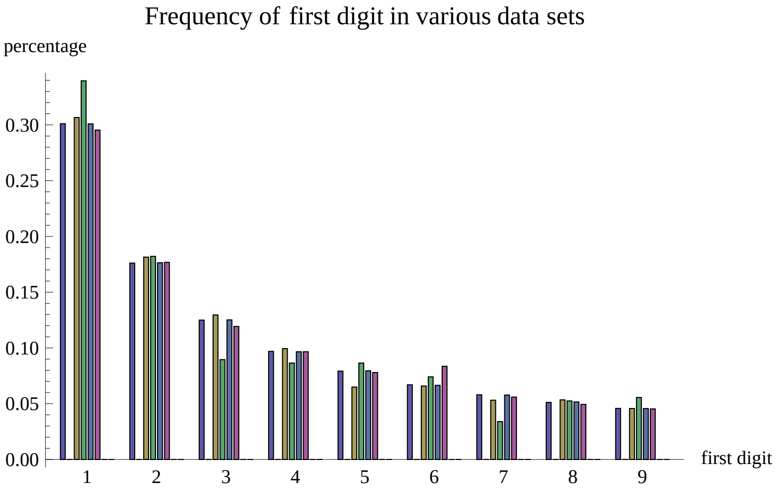


Figure 1.5 Frequencies of leading digits. The first column are the Benford probabilities, $\log_{10}\left(\frac{d+1}{d}\right)$, followed by (a) U.S. county populations (from 2000 census); (b) fundamental constants (from NIST); (c) first 3219 Fibonacci numbers; (d) first 3219 factorials. Note the census data includes Puerto Rico and the District of Columbia.

Definition 1.1.3 (Benford's Law for all the Leading Digits Base B (i.e., the significand))**.** *A data set satisfies Benford's Law for all the Leading Digits 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 summarizing what follows. The central questions we address are

1. `Why do so many data sets (mathematical expressions, physical data, financial transactions) follow this law?`
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. **PERHAPS LIST CHAPTER REFERENCES, PERHAPS SAY THE BOOK CAN SERVE AS AN INTRO TO SOME OF THESE. PERHAPS WAIT TILL END OF CHAPTER FOR DETAILS OR PERHAPS DO HERE AND AT END OF CHAPTER.**

The second question leads to many surprising characters entering the scene. The reason Benford's law is not just a curiousity 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 appear. Some of these laws are quite common, and thus are easily incorporated by people modifying data. For example, everyone knows that if you simulate flipping a fair coin 1,000,000 times then there should be about 500,000 heads. 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.

DESCRIBE THE REST OF THE CHAPTER. HISTORICAL INTRO, EXAMPLE, SUMMARY OF WHAT IS TO COME, ... PERHAPS WAIT TILL END OF CHAPTER FOR DETAILS OR PERHAPS DO HERE AND AT END OF CHAPTER.

1.2 HISTORY

In 1938 Frank Benford [**?**] published an article in the Proceedings of the American Philosophical Society titled *The law of anomalous numbers*. In it he 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, $\dots$), 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.

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 is typically closer to Benford's law. This is seen both in the row corresponding to n, $n^2, \dots$ (where we can show each of these is non-Benford) as well as in the average over all data sets. **TALK ABOUT TED HILL'S EXPLANATION.**

Though his article was enormously popular (at least for a scientific article), it was not the first paper on the subject. That honor goes to Simon Newcomb's *Note on the frequency of use of the different digits in natural numbers* in the American Journal of Mathematics in 1881. Unfortunately for Newcomb, his article [**?**] was a mere curiousity, and it was not until Benford's paper that other researchers were

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	54	1165
Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
$n, n^2, \dots n!$	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 1.1 Distribution of leading digits from the data sets Benford's paper [**?**]. Note the agreement with Benford law is best (and extremely good) for the amalgamated set of all data.

intrigued and began to study the subject. See Hurlimann's extensive bibliography [**?**] for a list of papers, books and reports on Benford's law from 1881 to 2006.

While our main theoretical emphasis in this book is to explain why and when Benford's law holds, it is worthwhile to take a few minutes to reflect on how it was discovered. 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, and the analysis of this leads to one of the most important theorems in probability and statistics.

Instead of looking at logarithm tables, we look at the steps in an old building. 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 than the cumulative effect will be visually apparent.

Below is a picture of the steps in **Nassau Hall** of Princeton University. **FIND A PICTURE OF THIS OR ANOTHER BUILDING!** Built in 1754, this building has seen a lot of history, from the Battle of Princeton in the Revolutionary War to serving as the capitol of the United States from July to October in 1783. One immediately sees that 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.

1.3 FUNDAMENTAL EQUIVALENCE

In mathematics, frequently progress is made by converting a problem to another one which has previously been analyzed and solved. While at first the probabilities in Benford's law seems surprising, after a simple change of variables we find that a data set $\{x_n\}$ is Benford if and only if $\{y_n := \log_{10} x_n\}$ is equidistributed modulo 1. To explain this equivalence, we first quickly review modular arithmetic. For more details, see [**?**] (or almost any book on group theory or abstract algebra).

Definition 1.3.1 (Modular (or clock) arithmetic)**.** *We say* $a \equiv b \bmod n$ *if* $a - b$ *is a multiple of* n*. This is frequently called clock arithmetic, as this is the most common example; on a clock, 13 o'clock and 1 o'clock are both represented by 1. We write* $y \bmod 1$ *for the unique real number* $r \in [0, 1)$ *such that* $y \equiv r \bmod 1$.

Definition 1.3.2 (Equidistributed modulo 1)**.** *A sequence* $\{z_n\}_{n=-\infty}^{\infty}$ *is equidistributed modulo 1 if*

$$\lim_{N\to\infty} \frac{\#\{n : |n| \le N, z_n \bmod 1 \in [a,b]\}}{2N+1} \;=\; b-a \tag{1.1}$$

for all $[a,b] \subset [0,1]$*. A similar definition holds for* $\{z_n\}_{n=0}^{\infty}$*.*

This equivalence is at the heart of many investigations on Benford's law. Equidistributed sequences have been extensively studied; reformulating the problem in this language allows us to use these tools and results for our problems. We give a quick sketch of the proof of this equivalence; applying this equivalence will occupy a large part of the book.

Lemma 1.3.3. *The significand (i.e., all the leading digits) of* 10^u *and* 10^v *are the same if and only if* $u \equiv v$ *mod* 1*.*

Proof. We prove one direction as the other is similar. If $u \equiv v$ mod 1, we may write $v = u + m$, $m \in \mathbb{Z}$. If

$$10^u \;=\; u_k 10^k + u_{k-1}10^{k-1} + \cdots + u_0 + u_{-1}10^{-1} + \cdots\,, \tag{1.2}$$

then

$$\begin{aligned} 10^v \;&=\; 10^{u+m} \\ &=\; 10^u \cdot 10^m \\ &=\; (u_k 10^k + u_{k-1}10^{k-1} + \cdots + u_0 + u_{-1}10^{-1} + \cdots)10^m \\ &=\; u_k 10^{k+m} + \cdots + u_0 10^m + u_{-1}10^{m-1} + \cdots\,. \end{aligned} \tag{1.3}$$

he first digit of each is u_k, the second digit of each is u_{k-1} and so on, which shows the two numbers have the same significand. □

We can use the above lemma to construct a simple test of whether or not two numbers have the same significand (i.e., the same leading digits). We first set some notation. Consider the unit interval $[0,1)$. For $d \in \{1,\dots,9\}$, define p_d by

$$10^{p_d} = d \quad \text{or equivalently} \quad p_d \;=\; \log_{10} d. \tag{1.4}$$

For $d \in \{1,\dots,9\}$, let

$$I_d \;=\; [p_d, p_{d+1}) \;\subset [0,1). \tag{1.5}$$

Lemma 1.3.4. *The first digit of* 10^y *is* d *if and only if* $y \bmod 1 \in I_d$*.*

Proof. By Lemma 1.3.3 we may assume $y \in [0,1)$. Then $y \;\in I_d = [p_d, p_{d+1})$ if and only if $10^{p_d} \;\le\; y \;<\; 10^{p_{d+1}}$, which from the definition of p_d is equivalent to $d \le 10^y < d+1$, proving the claim. □

The content of the above lemmas is that studying all the leading digits of a sequence $\{x_n\}$ is equivalent to studying the sequence $\{y_n\}$ modulo 1, where $y_n = \log_{10} x_n$. In other words, the distribution of the significands of $\{x_n\}$ can be determined by the distribution of the mantissas of $\{y_n\}$.

The following theorem shows that the exponentials of equidistributed sequences are Benford, and yields one of the most powerful and frequently used tests to prove Benford behavior.

Theorem 1.3.5 (Fundamental Equivalence for Benford's Law)**.** *A sequence $y_n = \log_{10} x_n$ is equidistributed modulo 1 if and only if the sequence x_n satisfies Benford's Law.*

Proof. We give the proof in base 10; the proof for a general base follows similarly.

As a warm-up, we first show how $\{y_n\}$ being equidistributed implies the first digit of $\{x_n\}$ satisfies Benford's Law. By Lemma 1.3.4,

$$\begin{aligned}&\{n \le N : y_n \bmod 1 \in [\log_{10} d, \log_{10}(d+1))\}\\ &\quad = \{n \le N : \text{first digit of } x_n \text{ is } d\}\,. \end{aligned} \tag{1.6}$$

Therefore

$$\begin{aligned}&\lim_{N\to\infty} \frac{\#\{n \le N : y_n \bmod 1 \in [\log_{10} d, \log_{10}(d+1))\}}{N}\\ &\quad = \lim_{N\to\infty} \frac{\#\{n \le N : \text{first digit of } x_n \text{ is } d\}}{N}. \end{aligned} \tag{1.7}$$

If y_n is equidistributed, then the left side of (1.7) is $\log_{10}\left(\frac{d+1}{d}\right)$ which implies x_n is Benford.

The general argument is almost identical. We wish to show $\{y_n\}$ is equidistributed if and only if the significands of $\{x_n\}$ satisfy Benford's Law. Note that the probability that $y_n \in [a,b) \subset [0,1)$ is just the probability that $y_n \in [0,b)$ minus the probability that $y_n \in [0,a)$. Thus, without loss of generality, the equidistribution of $\{y_n\}$ follows from showing that the probability that $y_n \in [0,t)$ equals t for any t. Let $s \in [1,10)$ be any significand, so $\log_{10} s \in [0,1)$. We have

$$\begin{aligned}&\lim_{N\to\infty} \frac{\#\{n \le N : y_n \bmod 1 \in [0, \log_{10} s)\}}{N}\\ &\quad = \lim_{N\to\infty} \frac{\#\{n \le N : \text{significand of } x_n \text{ is } s\}}{N}. \end{aligned} \tag{1.8}$$

If $\{y_n\}$ is equidistributed then the left hand side above is $\log_{10} s$, implying that $\{x_n\}$ is Benford. Conversely, if $\{x_n\}$ is Benford then the right hand side is $\log_{10} s$, implying that the probability of y_n being at most $\log_{10} s$ is $\log_{10} s$, which from our discussion above proves that $\{y_n\}$ is equidistributed. □

Remark 1.3.6. While we want an equivalence between a data set being Benford and its logarithm modulo 1 being equidistributed, we cannot have such an equivalence if we restrict ourselves to studying just the first digit in one base. To see this, imagine a sequence $\{x_n\}$ such that $\log_{10} x_n \bmod 1\{p_1, \dots, p_9\}$. While it is possible to choose the sequence so that it is Benford and its logarithm obeys the above restriction, such a sequence cannot be equidistributed as it only takes on 9 values.

This is one reason why we extended the definition of Benford's Law from a statement concerning the distribution of the first digit to the distribution of the entire significand. Our theorem then states $\log_B x_n \bmod 1$ is equidistributed if and only if x_n is Benford base B, which is the **fundamental equivalence** in the subject. See [**?**] for details.

DECIDE WHETHER OR NOT TO HAVE EXERCISES

Exercise 1.3.7 (Basic Properties of Congruences)**.** *Prove the following: For a fixed positive integer n and a, a', b, b' integers we have*

1. $a \equiv b \bmod n$ *if and only if* $b \equiv a \bmod n$.
2. $a \equiv b \bmod n$ *and* $b \equiv c \bmod n$ *implies* $a \equiv c \bmod n$.
3. $a \equiv a' \bmod n$ *and* $b \equiv b' \bmod n$*, then* $ab \equiv a'b' \bmod n$. *In particular* $a \equiv a' \bmod n$ *implies* $ab \equiv a'b \bmod n$ *for all* b.
4. *If* $a \equiv a' \bmod n$*, then for any polynomial* f *with integer coefficients we have* $f(a) \equiv f(a') \bmod n$.

Exercise 1.3.8 (Advanced)**.** *Define multiplication of* $x, y \in \mathbb{Z}/n\mathbb{Z} := \{0, 1, \dots, n-1\}$ *by* $x \cdot y$ *is the unique* $z \in \mathbb{Z}/n\mathbb{Z}$ *such that* $xy \equiv z \bmod n$. *We often write* xy *for* $x \cdot y$. *Prove that this multiplication is well defined. Using the Euclidean algorithm, prove that an element* x *has a multiplicative inverse if and only if* x *and* n *are relatively prime. Conclude that if every non-zero element of* $\mathbb{Z}/n\mathbb{Z}$ *has a multiplicative inverse, then* n *must be prime.*

Exercise 1.3.9 (Divisibility Rules)**.** *Prove a number is divisible by 3 (or 9) if and only if the sum of its digits are divisible by 3 (or 9). Prove a number is divisible by 11 if and only if the alternating sum of its digits is divisible by 11 (for example, 341 yields 3-4+1). Find a rule for divisibility by 7.*

Exercise 1.3.10. *If* $ab \equiv ac \bmod n$ *must* $b \equiv c \bmod n$*?*

Exercise 1.3.11. *Prove the other direction of the if and only if in Lemma 1.3.3.*

1.4 EXAMPLE

In Table 1.4 we list the **Benford probabilities** for the first digit and the observed probabilities for the first digits of the sequence 2^n, where in the first case we study all $n \in \{1, 2, \dots, 60\}$ while in the second we study all $n \in \{1, 2, \dots, 60000\}$.

A detailed analysis of this example illuminates many of the key features and challenges in developing the theory and applications of Benford's law. We have seen in §1.3 that a sequence $\{x_n\}$ is Benford if and only if the sequence $\{y_n := \log_{10} x_n\}$ is equidistributed modulo 1. We have thus reduced our problem to determining when the logarithms modulo 1 are equidistributed. In many cases, this follows from a well-known result in number theory or ergodic theory; see Theorem 12.3.5, Chapter 12 of [**?**] for a proof.

Theorem 1.4.1 (Kronecker-Weyl)**.** *Let* α *be an irrational number. Then the sequence* $z_n = n\alpha$ *is equidistributed modulo 1.*

Digit	Benford probabilities	2^n for $n \leq 60$	2^n for $n \leq 60000$
1	30.103%	30.000%	30.102%
2	17.609%	20.000%	17.610%
3	12.494%	10.000%	12.493%
4	9.691%	10.000%	9.692%
5	7.918%	10.000%	7.918%
6	6.695%	6.667%	6.695%
7	5.799%	3.333%	5.798%
8	5.115%	8.333%	5.117%
9	4.576%	1.667%	4.575%

Table 1.2 The Benford probabilities and the observed probabilities for the first digit frequencies of the first 60 and the first 60000 numbers of the form 2^n.

As $\log_{10} 2$ is irrational, $y_n = \log_{10} 2^n = n \log_{10} 2$ is thus equidistributed modulo 1, and therefore the sequence $x_n = 2^n$ is Benford.

Appealing to this advanced machinery, we see that if we take sufficiently many terms then we will observe Benford behavior in the sequence 2^n. This raises the natural question: how many terms suffice? From an applications point of view, this is extremely important. For example, the IRS wants to (and does!) use Benford's law to determine probable tax fraud. It is not enough to tell them to audit any corporation whose distribution of leading digits of numbers in tax returns is not close to Benford whenever there are *sufficiently many* data points in the return unless we also tell them what how large *sufficiently many* is.

Looking at the data in Table 1.4, it is not clear if 60 terms is sufficient. Though the probabilities are quite good for some digits (such as 1, 2, 3 and 4), the probabilities seem a bit off for others (especially 9). Is this just a result of having a small data set, or will these patterns persist?

To see which is the case requires a statistical analysis of the data, which will test how likely such an observation is under the assumption that the data follows Benford's law. The correct test here is a **chi-square goodness of fit test**. (REFER TO AN APPENDIX FOR DETAILS OF THIS TEST) For the first 60 terms of 2^n, we find a chi-square statistic of 3.78, significantly less than the 15.51 value for the 95% confidence interval. Thus the fit with Benford is significant. Further, if we increase the size of our data set to study the first 60,000 values of 2^n, we get a ridiculously low chi-square value of .0005, indicating phenomenal agreement with Benford's law. These results are not surprising, as the advanced number theory or ergodic machinery referred to above allows us to prove that 2^n is Benford.

Though the data for 2^n is highly suggestive of rapid convergence to the Benford probabilities, it turns out that it *is* significant that we only observe one number whose first digit is a 9, namely $9007199254740992 = 2^{53}$. The reason becomes apparent when we look more closely at the data. In Table 1.4 we list the first 60 values of 2^n ($n \in \{0, \dots, 59\}$).

1	1024	1048576	1073741824	$1.10\ldots \cdot 10^{12}$	$1.13\ldots \cdot 10^{15}$
2	2048	2097152	2147483648	$2.20\ldots \cdot 10^{12}$	$2.25\ldots \cdot 10^{15}$
4	4096	4194304	4294967296	$4.40\ldots \cdot 10^{12}$	$4.50\ldots \cdot 10^{15}$
8	8192	8388608	8589934592	$8.80\ldots \cdot 10^{12}$	$9.01\ldots \cdot 10^{15}$
16	16384	16777216	17179869184	$1.76\ldots \cdot 10^{13}$	$1.80\ldots \cdot 10^{16}$
32	32768	33554432	34359738368	$3.52\ldots \cdot 10^{13}$	$3.60\ldots \cdot 10^{16}$
64	65536	67108864	68719476736	$7.04\ldots \cdot 10^{13}$	$7.21\ldots \cdot 10^{16}$
128	131072	134217728	$1.37\ldots \cdot 10^{11}$	$1.41\ldots \cdot 10^{14}$	$1.44\ldots \cdot 10^{17}$
256	262144	268435456	$2.75\ldots \cdot 10^{11}$	$2.81\ldots \cdot 10^{14}$	$2.88\ldots \cdot 10^{17}$
512	524288	536870912	$5.50\ldots \cdot 10^{11}$	$5.63\ldots \cdot 10^{14}$	$5.76\ldots \cdot 10^{17}$

Table 1.3 The first 60 numbers of the form 2^n ($n \in \{0, \dots, 59\}$. As we are only interested in the leading digit, for the larger numbers we only display the first three digits in the mantissa.

The data is deliberately presented in this manner, with six columns of ten entries each. Notice the remarkable consistency as we move across a row. For example, the leading digit of every entry in the first row is always a 1; in fact, the leading digit is constant in rows 1, 2, 3, 5, 6, 8, 9 and 10. In the fourth row it is almost always an 8, but the last entry just barely is recorded as a 9, as it is $9007199254740992 = 9.007 \cdot 10^{15}$. This is no coincidence, but a direct consequence of the fact that $2^{10} = 1024 \approx 10^3$. Thus we see that the leading digits of 2^n are *almost* periodic with period 10; every 10 iterations our mantissa is increased by 2.4%. This is quite interesting – although initially there are a dearth of 9s, in the limit they will occur with the correct frequency. **MENTION THAT LATER IN THE BOOK WE'LL TALK ABOUT STUFF LIKE THIS IN GREATER DETAIL, LOOK AT PERIODS IN PI ETCETERA.**

This example illustrates that there can be small patterns that are missed by statistical tests. In the case of 2^n, it is due to an almost periodic nature. In general, questions about the rate of convergence are difficult and subtle. For 2^n, the answer is related to computing the error term in the Kronecker-Weyl Theorem. One approach involves the **Erdös-Turan Theorem**, which gives the needed convergence rates as a function of how well α is approximated by rationals. **(ADD REFERENCE TO THIS IN THE BOOK)**

Remark 1.4.2. The fact that $2^{10} \approx 10^3$ is well known to computer scientists; in fact, a gigabyte is not (as the name would suggest) 1000 megabytes, but rather 1024.

MUST DECIDE WHETHER OR NOT WE WANT EXERCISES.

Exercise 1.4.3. *Show the Benford probabilities* $\log_{10}\left(\frac{d+1}{d}\right)$ *for* $d \in \{1, \dots, 9\}$ *are irrational. What if instead of base ten we work in base* B *for some integer* B*? What would the corresponding probabilities equal, and are they irrational?*

Exercise 1.4.4. *Prove that* $\log_{10} 2$ *is irrational.* Hint: assume not; then it must equal some rational p/q with p and q relatively prime.

Exercise 1.4.5. *Note $2^{50} = 1.13 \cdot 10^{15}$ and $2^{53} = 9.01 \cdot 10^{15}$. Find the smallest $m \geq 0$ such that the leading digit of 2^{50+10m} is not a 1, and the smallest m' such that the leading digit of $2^{53+m'}$ is not a 9. Is it surprising that $m > m'$?*

Exercise 1.4.6. *As $N \to \infty$, do you expect the probability of observing a 9 as the first digit of 2^n for $n \leq N$ to more often be greater than or more often to be less than $\log_1 0 \left(\frac{10}{9}\right)$?*

Exercise 1.4.7. *Prove $n\alpha \bmod 1$ cannot be equidistributed if α is a rational number.*

Exercise 1.4.8 (Fibonacci numbers)**.** *The **Fibonacci numbers** are defined by the recurrence relation $F_{n+2} = F_{n+1} + F_n$, with initial conditions $F_0 = 0$ and $F_1 = 1$. Show that this uniquely determines F_m for all m, and show that $F_n = \frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^n - \frac{1}{\sqrt{5}}\left(\frac{1-\sqrt{5}}{2}\right)^n$. With some careful book-keeping and appealing to the Kronecker-Weyl Theorem, show this implies the Fibonacci numbers are Benford.* ***REFERENCE LATER PART IN BOOK ON DIFFERENCE EQS AND BENFORD'S LAW, AS WELL AS MILLER - TAKLOO-BIGHASH.***

Exercise 1.4.9. *Let $x_n = ar^n$ with $\log_{10} r$ irrational. Prove x_n is Benford.*

Exercise 1.4.10. *Do the first digits of e^n follow Benford's Law? What about $e^n + e^{-n}$?*

Exercise 1.4.11. *Give an example of a sequence $\{x_n\}$ and two integers $b_1, b_2 > 1$ such that $\{x_n\}$ is Benford base b_1 but* not *Benford base b_2.*

1.5 EXPLANATIONS FOR BENFORD'S LAW

Now that we have stated Benford's Law and seen examples of data sets that are close to it as well as others that significantly deviate, it is worth giving some informal arguments and heuristics as to when we should expect a data set to satisfy Benford's Law. We will flush these arguments out in greater detail later in the book; the remarks below are meant to introduce the material that follows. We describe three approaches which lead to Benford behavior, and comment on how other arguments fall into these categories: (1) being spread out over several orders of magnitude, (2) products of random variables and the Central Limit Theorem, and (3) scale invariance. As there are many common features between (1) and (2), we begin with a somewhat detailed review of probability concepts, which will be useful in the formulations and arguments that follow.

1.5.1 Spread out distributions

In Theorem 1.3.5 we noted that a sequence $\{x_n\}$ is Benford if and only if $\{y_n := \log_{10} x_n\}$ is equidistributed modulo 1. In other words, the signficands of $\{x_n\}$ follow Benford's law if and only if the fractional parts of the logarithms (the mantissas) fall uniformly in $[0, 1]$. This allows us to reformulate our investigation into a study of when logarithms are equidistributed modulo 1.

In his classic text *An Introduction to Probability Theory and Its Applications* (Volume II), Feller [**?**] stated that if a random variable X is spread out over many orders of magnitude, then its logarithm modulo 1 will be approximately equidistributed, and we will obtain Bendford behavior. This viewpoint is expanded by Fewster [**?**]. While one can of course construct counter-examples to this intuition (we give one at the end of this section), the hope is that this should be a simple test and a reasonable explanation for the prevalence of Benford Behavior. We briefly summarize their arguments.

Let X be a continuous random variable with **probability density function** $f_X(x)$ (the subscript X is to remind us that this is the density for the random variable X). This means that the probability $X \in [a,b]$ is just $\int_a^b f_X(x)dx$, and of course $\int_{-\infty}^{\infty} f_X(x)dx = 1$. The **cumulative distribution function** of X is the probability that X is at most x, or $\text{Prob}(X \le x)$. If we let F_X denote the anti-derivative of f_X with $\lim_{x\to-\infty} F_X(x) = 0$, then the cumulative distribution function at x is just $F_X(x)$. This follows immediately from the definition, as the probability X is at most x is

$$\text{Prob}(X \le x) \ = \ \int_{-\infty}^{x} f_X(x)dx \ = \ F_X(x) - F_X(-\infty) \ = \ F_X(x). \tag{1.9}$$

By the Fundamental Theorem of Calculus, the derivative of the cumulative distribution function is the density; this is a very convenient way to compute probability densities.

For convenience, we assume $f_X(x) = 0$ if $x \le 0$ so that X only takes on positive values. Given a continuous random variable X with density f_X, what is the density of the random variable $Y = \log_{10} X$? We can readily compute the cumulative distribution function of Y; it is

$$\text{Prob}(Y \le y) \ = \ \text{Prob}(\log_{10} X \le y) \ = \ \text{Prob}(X \le 10^y) \ = \ F_X(10^y), \tag{1.10}$$

where the last equality follows from the fact that F_X gives the cumulative distribution function of X. To find the density of Y we simply differentiate both sides. Denoting this density by f_Y, we have

$$f_Y(y) \ = \ \frac{d}{dy}F_X(10^y) \ = \ F_X'(10^y)\frac{d10^y}{dy} \ = \ f_X(10^y)10^y \log 10; \tag{1.11}$$

this follows from the Chain Rule, and noting that the derivative of F_X is f_X.

The probability that $Y \bmod 1$ is in $[a,b] \subset [0,1]$ is the union of the probabilities that $Y \in [a+n, b+n]$ for some integer n. This is just

$$\begin{aligned}\text{Prob}(Y \bmod 1 \in [a,b]) \ &= \ \sum_{n=-\infty}^{\infty} \int_{a+n}^{b+n} f_Y(y)dy \\ &= \ \sum_{n=-\infty}^{\infty} \int_{a+n}^{b+n} f_X(10^y)10^y \log 10 dy. \end{aligned} \tag{1.12}$$

Alternatively, we may view this as

$$\text{Prob}(Y \bmod 1 \in [a,b]) \ = \ \int_a^b g_Y(y)dy, \tag{1.13}$$

where

$$g_Y(y) \;=\; \sum_{n=-\infty}^{\infty} f_Y(y+n) \;=\; \sum_{n=-\infty}^{\infty} f_X(10^{y+n})10^{y+n}\log 10. \tag{1.14}$$

Feller and Fewster argue that if X is spread over several orders of magnitude, the averaging process used to construct the density g_Y of $\log_{10} X \bmod 1$ yields a density which is close to the uniform density on $[0,1]$; note that if the density of $Y \bmod 1$ is the uniform density on $[0,1]$ then Y is equidistributed modulo 1, as the probability of $Y \bmod 1 \in [a,b]$ would just be $\int_a^b 1dx = b-a$ as required. In later chapters we will make these arguments precise for many distributions by using Poisson Summation, which converts the slowly converging sums over n of f_X to rapidly converging sums over n of $\widehat{f_X}$, the Fourier transform of f_X.

It is informative to revisit our example of U.S. county populations and the NYSE trading volumes. In Figure 1.3 we see that the county populations are spread out over several orders of magnitude, while the trading volumes are clustered together. This is even more apparent in Figure 1.4, where we see the county population has sizable probabilities over 3 or 4 units on a log-scale, while the stock volumes do not even span one unit. Thus, the spread-out hypothesis asserts that the county populations should be close to Benford's Law while the stock volume data should not. This is born out in Figure 1.1.

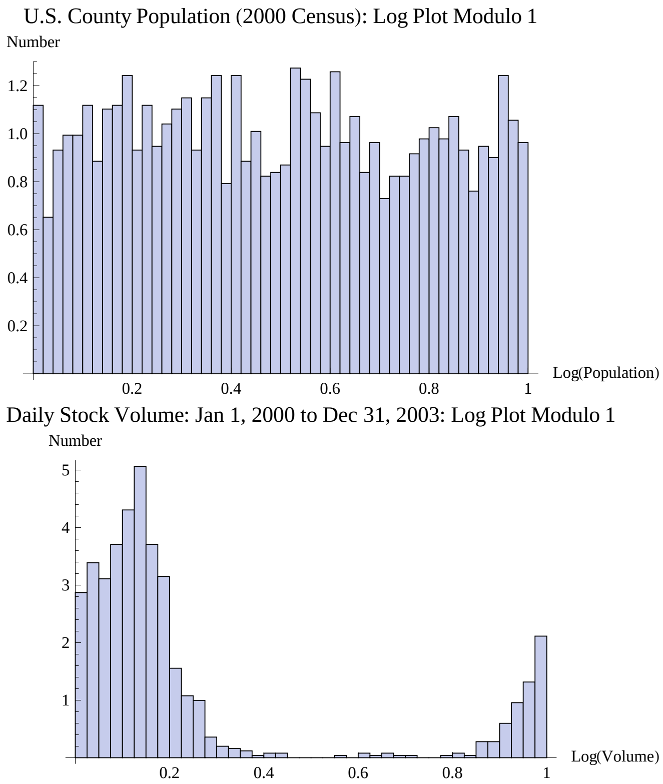
In the spirit of our arguments above, it is worthwhile to look at the distribution of the logarithms modulo 1 for these two examples. We plot these in Figure 1.6. As expected, logarithms modulo 1 of the county populations are close to the uniform density, while those of the stock volumes are decidedly not. For the stock volumes, it isn't surprising that most of the probability of the logarithms modulo 1 is concentrated near 0 and 1 and that the probability is low in the middle. Looking at the original data (the stock volumes) in Figure 1.3, we see that most of the amounts are clustered near 1 billion, with most of the amounts between 1 and 2.5 billion but a significant number just a little less than 1 billion. The effect of this, upon taking logarithms modulo 1, will be to have most of the probability clustered near 0 and 1, as the values from 10^9 to $2.5 \cdot 10^9$ have logarithms modulo 1 in $[0, .4]$ while values from $7 \cdot 10^8$ to 10^9 have logarithms modulo 1 lying in $[.85, 1]$.

We are thus led to our first explanation for the prevalence of Benford's Law.

Spread-out Heuristic for Benford Behavior: Any random variable whose values are spread out over several orders of magnitude should be close to satisfying Benford's Law.

Example 1.5.1. *As promised, we provide an example of a spread out distribution which is not Benford. Let*

$$f_X(x) \;=\; \begin{cases} \frac{6}{\pi^2 n} & \text{if } 10^n \le x \le 10^n + \frac{1}{n} \text{ for } n \text{ a positive integer} \\ 0 & \text{otherwise.} \end{cases} \tag{1.15}$$


U.S. County Population (2000 Census): Log Plot Modulo 1

Daily Stock Volume: Jan 1, 2000 to Dec 31, 2003: Log Plot Modulo 1

Figure 1.6 (Top) The population of U.S. counties. (Bottom) The daily volume of the NYSE from 2000 through 2003.

Clearly $f_X(x)$ is non-negative, and

$$\int_{-\infty}^{\infty} f_X(x)dx \ = \ \sum_{n=1}^{\infty} \int_{10^n}^{10^n+\frac{1}{n}} \frac{6}{\pi^2 n} dx \ = \ \frac{6}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2} \ = \ 1 \qquad (1.16)$$

*as $\sum_{n=1}^{\infty} 1/n^2 = \pi^2/6$ (see §11.3.4 of [**?**], or almost any book on Fourier Series). Thus f_X is a probability density, and clearly the first digit of X is always 1. Note this distribution is extremely spread out; in fact, both the mean and variance are infinite!*

1.5.2 Products of Random Variables

Many quantities in the real world are obtained by multiplying different quantities (or performing even more elaborate operations to combine them). For example, consider the price of a gold bar whose dimensions are randomly chosen. The value of the bar equals $\rho xyzd$, where ρ is the density, x, y and z are the dimensions and d is the cost of gold per gram. For another example, consider the ideal gas law from Chemistry / Physics, which tells us that the pressure P equals nRT/V with n the number of moles, R a universal constant, T the temperature and V the volume. For one final example, we turn to the equation of exchange in economics, which asserts that the amount of money circulating on average equals PQ/V with P the price level, Q the index of expenditures and V the velocity of money (i.e., how frequently it is spent).

Thus understanding the distribution of digits of these derived quantities requires us to understand the effect of multiplying different random variables. Numerous authors (see for example Hamming [**?**]) observed that the product of two random variables is typically closer to satisfying Benford's Law than either, and the more products taken the closer the behavior is to Benford. This leads to our second heuristic. We provide an explanation as to why this is the case in a simple situation where we can use the Central Limit Theorem; later in the book we discuss how to handle the general case.

Before stating the Central Limit Theorem, we recall the set-up. Assume $W_1, W_2, \dots$ are independent identically distributed random variables (i.i.d.r.v.) drawn from a density p with mean μ and variance σ^2; thus, $\text{Prob}(X_i \in [a, b]) = \int_a^b p(x)dx$. Recall the density of a normal (or Gaussian) random variable with mean μ and variance σ is $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$. We state the simplest case of the Central Limit Theorem, which can be generalized in certain cases to apply to sums of non-identically distributed random variables.

Theorem 1.5.2 (Central Limit Theorem)**.** *Let W_i be as above and assume the third moment of each W_i is finite (i.e., $\int_{-\infty}^{\infty} |x|^3 p(x)dx < \infty$). If $S_N = W_1 + \cdots + W_N$, then S_N converges in probability to being normally distributed with mean $N\mu$ and variance $N\sigma^2$; equivalently,*

$$\lim_{N\to\infty} \text{Prob}\left(\frac{S_N - N\mu}{\sigma\sqrt{N}} \ \in \ [a, b]\right) \ = \ \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx. \qquad (1.17)$$

Theorem 1.3.5 is our main tool for proving Benford's Law; it asserts that a data set is Benford if and only if its logarithms are equidistributed modulo 1. In other words, X is Benford if and only if $Y = \log_{10} X$ is equidistributed modulo 1. Imagine now that X is a product of many quantities, say $X = X_1 \cdots X_N$. Taking logarithms we find

$$Y \ = \ \log_{10} X \ = \ \log_{10}(X_1 \cdots X_N) \ = \ \log_{10} X_1 + \cdots + \log_{10} X_N. \quad (1.18)$$

Letting

$$Y_i \ = \ \log_{10} X_i, \quad (1.19)$$

we find

$$Y \ = \ Y_1 + \cdots + Y_N. \quad (1.20)$$

If the X_i's are identically distributed random variables, then so too are the Y_i's, say with mean μ and variance σ^2. Hence by the Central Limit Theorem we have Y tending to being normally distributed with mean $N\mu$ and variance $N\sigma^2$. It is now a nice exercise in integration (which is simplified if one knows the Poisson Summation formula) to show that if Y is normally distributed with mean μ_Y and variance σ_Y^2, then as $\sigma_Y \to \infty$ we have $Y \bmod 1$ tends to the uniform distribution, which implies $X = 10^Y$ obeys Benford's Law. We will prove this in **ADD REFERENCE TO LATER CHAPTER**; see also Theorem 9.4.4 of [**?**].

If the distributions of the X_i's are nice then Y rapidly converges to being normally distributed with a large variance, which then becomes approximately uniformly distributed when we study it modulo 1. We demonstrate this rapidity by studying three Gaussians below. All have a mean of 4 and the standard deviations are .25, .5 and 1; see Figure 1.7 for the plots.

Taking the densities of he two Gaussians with standard deviations of .5 and 1 modulo 1 yields densities that are very close to the uniform distribution. To see the difference, we zoom in on these two in Figure 1.8.

This rapid convergence of Gaussians modulo 1 to the uniform distribution explains why numerous systems exhibit Benford behavior. Interpreting our problem along these lines, we see the connection between Benford's Law and the Central Limit Theorem. Formally, we isolate our second heuristic below.

Products of Random Variables Heuristic for Benford Behavior: Any random variable obtained by multiplying many independent quantities should be close to satisfying Benford's Law. Further, the more quantities are involved the closer it will be to Benford behavior. Similar results should hold for more general combinations than simple products.

In our first heuristic, the more spread out a generic data set was the closer we expected its behavior to be towards Benford's Law. We could see this by looking at the associated densities and noting that the summation over n was equivalent to an averaging. In the situation above, the analogue is the number of products; the more products (or the more involved operations we have), the closer we should be

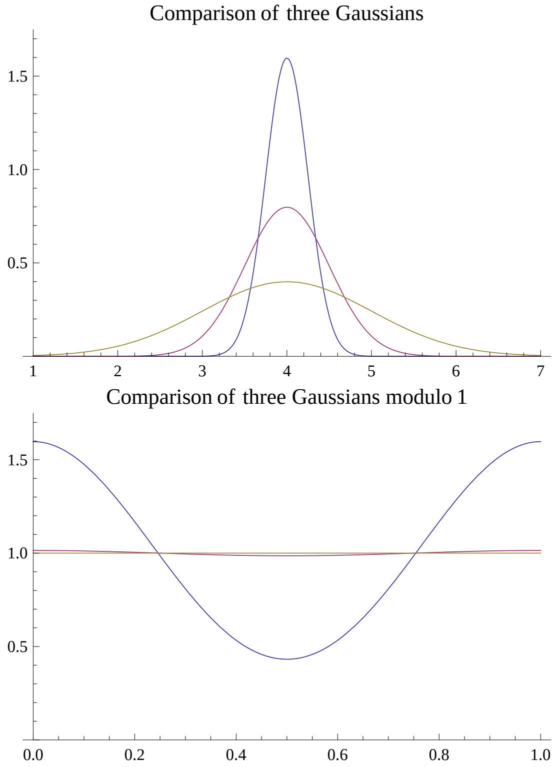


Figure 1.7 Plots of three Gaussian distributions. All have a mean of 4; the one with the highest peak has a standard deviation of .25, the next has a standard deviation of .5 and the one with the lowest peak has a standard deviation of 1. The top plot are the three Gaussians, while the bottom plot are the probability densities of the three Gaussians modulo 1.

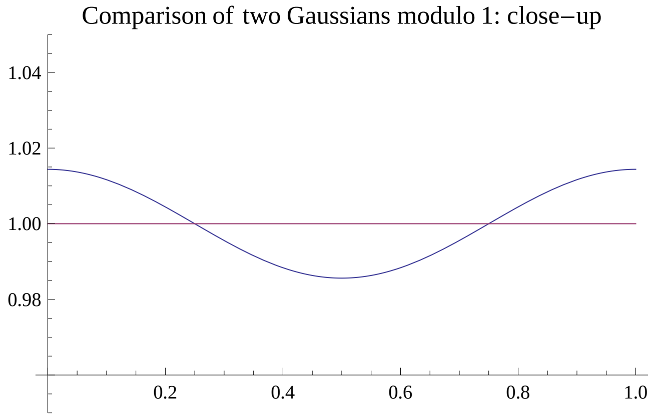


Figure 1.8 Plots of the density from two Gaussian distributions with mean 4 and standard deviations of .5 and 1 modulo 1. Note that already at a standard deviation of 1 the result is, to the eye, indistinguishable from the uniform distribution.

towards satisfying Benford's Law. When we make both of these heuristics precise below, we shall heavily draw on Fourier analysis, especially Poisson Summation. It is natural to expect these tools to enter. We have already seen how Fourier transforms and Poisson Summation are lurking in the spread out heuristic. For our second approach, recall that one standard method of proving the Central Limit Theorem is to use Fourier transforms. The reason is that the Fourier transform of a convolution is the product of the Fourier transforms. The **convolution** of two densities f and g is given by

$$(f * g)(x) \ = \ \int_{-\infty}^{\infty} f(t)g(x-t)dt. \tag{1.21}$$

The reason convolutions play such an important role in the subject is that if X and Y are random variables with densities f and g, then the density of the random variable $f + g$ is simply $f * g$. We will expand on these ideas in Chapter **ADD REFERENCE**.

1.5.3 Scale invariance

We conclude this section with yet another explanation for why so many data sets satisfy Benford's Law. Any real world data set includes not just numbers but also units. For example, we might be measuring pressure in pascals or the value of a stock in dollars. Clearly the first digit can depend on our choice of units. In some cases the significand is unchanged if we switch units; for example, the first digit of a stock is the same whether we measure it in dollars or cents; however, if we now measure it in euros, yens or in gold pressed latinum, we would obtain a different value. The same is true for pressure, where we could use pounds per square inch or inch of mercury.

Benford's Law is supposed to be a universal result for many sets of data. If this is to be the case, then the distribution of the first digit (or more generally the entire significand) must be independent of our choice of units. This property is called **scale invariance**. The only distribution that satisfies scale invariance is the Benford distribution, where the probability of a signficand of at most s base 10 is $\log_{10} s$.

We give a quick justification of this claim, and leave the complete argument to **ADD REFERENCE TO LATER IN THE BOOK.** Again by our Fundamental Equivalence (Theorem 1.3.5), a random variable X obeys Benford's Law if and only if $Y = \log_{10} X$ is equidistributed modulo 1 (or, equivalently, if $Y \bmod 1$ has the uniform distribution).

If the distribution of the significand is to be scale invariant, then then distribution of the signficands for X and $V = cX$ must be the same for any fixed, positive constant c. If we look at the logarithms, we have $Y = \log_{10} X$ and

$$W \ = \ \log_{10} V \ = \ \log_{10}(cV) \ = \ \log_{10} V + \log_{10} c \ = \ Y + \log_{10} c. \qquad (1.22)$$

We claim this forces $Y \bmod 1$ to be the uniform distribution. To see this, imagine that the density for $Y \bmod 1$ is not constant, taking on its maximum value at $y_{\max}$ and its minimum value at $y_{\min}$. For each of exposition, let's assume $y_{\max} = \log_{10} 6$ and $y_{\min} = \log_{10} 3$. We take $c = 10^{y_{\max}-y_{\min}} = 2$. This now shows that Y and W cannot have the same distribution, as the most likely value for Y is $\log_{10} 6$, which is the *least* likely value for W. Equivalently, X and V cannot have the same distribution for their signficands, as the probability density for the signficand of X is largest at 6, which is where the density is smallest for V.

Another way to see the above is to note that we require $Y \bmod 1$ and $W \bmod 1$ (where $W = Y + \log_{10} c$ to have the same distribution for any positive c. If the density of $Y \bmod 1$ were not constant, we choose our shift accordingly and see that the two densities do not align.

Returning to the 2000 census, we look at the land area of the counties in Figure 1.9. The first column is the Benford probabilities, followed by the observed frequencies for the first digit of the areas in miles^2, km^2 and acres. Not only is each data set non-Benford, but they all seem to follow different distributions.

For example, let's look at the land area in miles^2 (see Figure 1.10. Note that the data is clustered together; this leads to the logarithms modulo 1 deviating significantly from the uniform distribution.

On further reflection, it is perhaps not too surprising that the land areas of U.S. counties do not follow Benford's Law. Returning to colonial times, one can easily imagine scenarios where counties would be created to have a manageable size (see, for instance, the Northwest Ordinance and the ensuing development of the Midwest).

While land areas of counties created by man are not a good candidate for Benford's law, there is a related statistic that is an ideal candidate, namely population density. We compare the population densities of the counties in the various units in Figure 1.11.

Unlike Figure 1.9, the three scaled data sets in Figure 1.11 are much closer to having the same distribution for their leading digit (which is close to Benford's

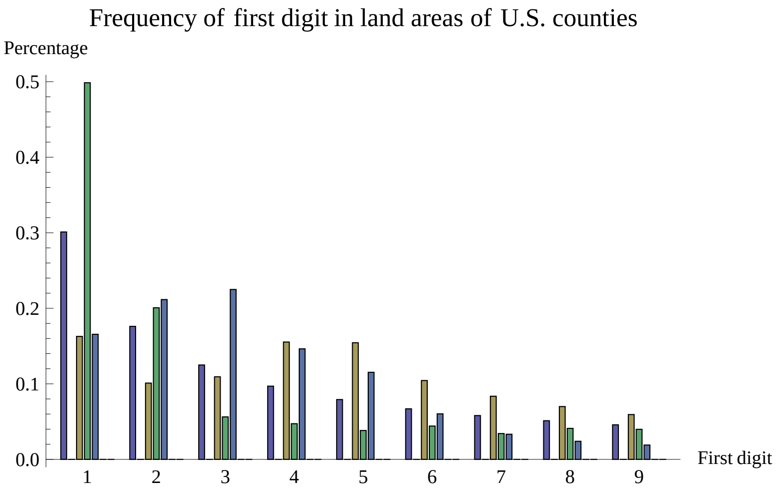


Figure 1.9 Plots of the first digit of the land area of U.S. counties from the 2000 census (including Puerto Rico and D.C.). The first column is the Benford probabilities, followed by the land area in miles^2, km^2 and acres.

Law). This leads us to our final heuristic to explain the universality of Benford behavior.

Scale Invariant Heuristic for Benford Behavior: The only distribution of leading digits which is invariant under rescaling the units is Benford's Law.

1.6 SUMMARY

IN THIS SECTION WE'LL SUMMARIZE WHERE BENFORD'S LAW OCCURS, DISCUSS THE VARIOUS TYPES OF PROOF TECHNIQUES (ESPECIALLY THEIR INGREDIENTS), AND MENTION THE APPLICATIONS AND GENERALIZATIONS.

1.6.1 Occurrences

1.6.2 Techniques

1.6.3 Applications

1.6.4 Generalizations

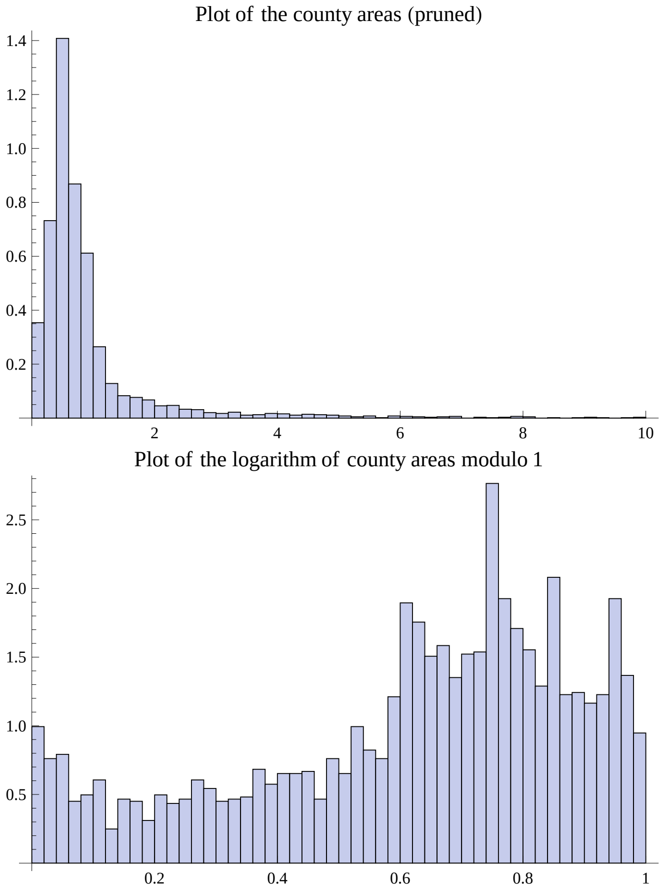


Figure 1.10 Plots of the first digit of the land area of U.S. counties in miles^2 from the 2000 census (including Puerto Rico and D.C.)

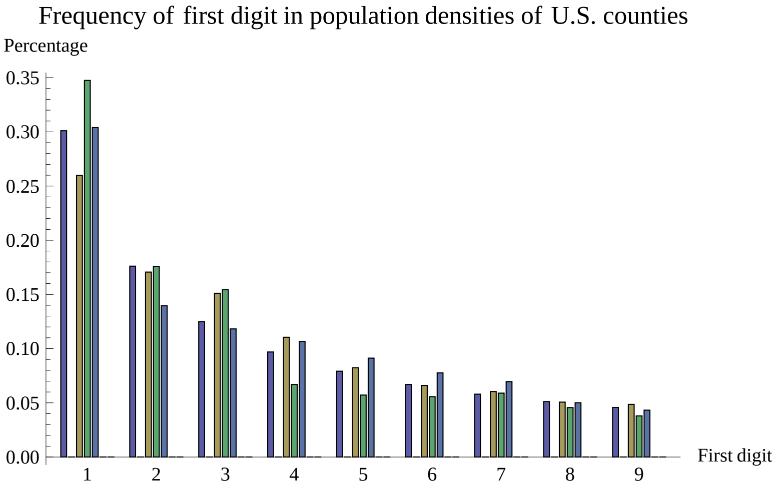


Figure 1.11 Plot of the population density of the counties in the U.S. from the 2000 census (including Puerto Rico). The first column is the Benford probabilities, followed by the population density using miles^2, km^2 and acres.