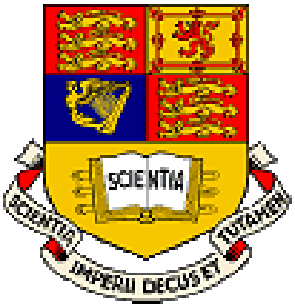
Imperial College of London
Department of Mathematics



Ecole Nationale Supérieure
d'Informatique et de
Mathématiques Appliquées de
Grenoble

ENSIMAG

Benford's Law

Adrien Jamain

April - September 2001.

Supervisor:
Prof. David J. Hand
(Imperial College)

Tutor:
Prof. Maryse Béguin
(Ensimag)

Contents

Acknowledgements

The author wishes to thank Prof. David Hand and Dr. Richard Bolton (Imperial College) for providing a constant and extensive support throughout the project, Prof. Theodore Hill (Georgia Institute of Technology) for communicating the apparently new result of section 2.2, and Dr. Pieter Allaart (University of North Texas) for explaining the proof of sum-invariance in section 2.4.

Introduction

Benford's Law is that kind of very counter-intuitive law which never ceases to astound. Actually some authors compared it with Newton's Law of gravitation, saying that it is more a simple observation of reality rather than a provable mathematical result. Dealing with that kind of law is quite difficult, because the scientist is then always somewhere between excessive cartesian doubt and mystical faith. However the author tried to stay the most realistic possible in his approach, without denying the reality of facts but without drawing any fast conclusion.

Benford's Law, after its discovery, was forgotten, or rather little studied, because it had little application to real problems. Nowadays, however, it is experiencing a regain of interest, certainly because of fraud detection and data mining cutting-edge applications. In that context the project was not aimed at those applications in the real world (some would say 'in the market'), but rather at the mathematical justifications of such applications.

The present report is a technical summary of some of the most interesting papers that were published on Benford's Law, plus a few experiments that the (skeptical) author performed on real datasets. The mathematical part includes nothing really new (except maybe section 2.2), as many of the theorems and the proofs were only detailed by the author, who thus made them easily understandable. Some of the sections are not as complete as the author would wish them to be, but all the major contributions are described.

Notations

The following notations and definitions will be used throughout this report:

- $\log(x)$ will denote the logarithm of x in base 10, while $\ln(x)$ will denote the natural logarithm of x (i.e. in base e), and $\log_b(x)$ the logarithm of x in base b.
- $\Pr(E)$ will denote the probability of the event E, and 1_E will denote the indicator function of E
- $[a, b]$ is the closed real interval, $[a, b)$ is the real interval closed in a and open in b, and similarly for $(a, b]$ and (a, b).
- If S is a set of numbers, k a real number, and n a positive integer, kS (or $k \times S$), S^k, and $S \mod n$ will denote respectively the sets $\{ks, s \in S\}$, $\{s^k, s \in S\}$, and $\{s \mod n, s \in S\}$.
- The cardinality of a countable set S will be denoted #S.
- The mantissa function in base b of a positive real number is defined as follow:

$$\begin{aligned} M^b \quad : \quad & \mathbb{R}^+ \to [1, b) \\ & x \quad \mapsto \; m_b \end{aligned}$$

  where m_b is the unique number in $[1, b)$ such that $x = m_b \times b^n$ for some $n \in \mathbb{Z}$.
- The mantissa function in base 10 will be denoted simply M.
- The k^{th} significant digit function in base b of a positive real number is defined as follow:

$$\begin{aligned} D_k^b \quad : \quad & \mathbb{R}^+ \to \{0, \ldots, b\} \\ & x \quad \mapsto \; d_k^b \end{aligned}$$

  where $\{d_k^b\}_k$ is the unique sequence s.t. $d_1^b \in \{1, \ldots, b-1\}, d_k^b \in \{0, \ldots, b-1\}$, and:

$$M^b(x) = \sum_{i=1}^{\infty} b^{-(i-1)} d_i^b$$

  .
- As for the mantissa, the base 10 will be omitted when possible in the notation of the significant digits.
- Unless stated otherwise, X will denote an arbitrary random variable, M the random variable representing its mantissa in base 10 ($M = M(X)$), and D_k the random variable representing its k^{th} significant digit in base 10 ($D_k = D_k(X)$).

Part I

Historical and Mathematical Background

Chapter 1

The discovery of Benford's law

This chapter is dedicated to the history of Benford's law, which is a good example of how discoveries are sometimes made, forgotten, and then found again...

1.1 1881: Newcomb's article

Simon Newcomb (1835-1909) is thought to have been the first to discover the phenomenon that would later be called Benford's law, or at least the first who published something about it. He was a highly honored American astronomer, his most famous work regarding planetary theories and astronomical constant derivation. His discovery of the departure of the moon from its predicted position led to the investigation of variations in rate of rotation of the earth.

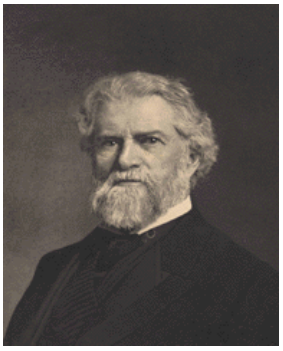


Figure 1.1: Simon Newcomb (1835-1909)

Simon Newcomb was not a mathematician in the strict sense of the term, but not surprisingly by its counter-intuitive nature Benford's law tends to attract more attention from physicists or physics related scientists than from pure mathematicians. In 1881, he published a two-pages long paper in the American Journal of Mathematics (see [Newcomb]), and he would surely be surprised how much the subject has been explored since then.

So what did Newcomb find?

> 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 ones wear out than the last ones.

According to his observations, the first pages in the logarithmic tables, i.e. those showing the logarithms of numbers beginning with the lowest digits, were more referred to than the last ones. His fellow scientists would thus be using more numbers beginning with the digit 1 than expected... This may sound strange at first sight, why should not the digits of such numbers be uniformly distributed? However, it makes more sense when one thinks that the numbers used by scientists are neither *purely random*, nor *purely deterministic*... In general these numbers come from some physical constants, by some derivation or some computation. Therefore Newcomb called them "natural number", maybe by default of a better name, whereas later Benford used the expression "outlaw numbers" in quite the same meaning.

Using some heuristic arguments about numbers, including some that might sound strange today ("As natural numbers occur in nature, they are to be considered as the ratios of quantities."), Newcomb concluded:

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

This sentence implies (although not clearly, see 2.1.4) that the probabilty that the first digit D_1 equals d is given by:

$$\Pr(D_1 = d) = log\left(1 + \frac{1}{d}\right)$$

However, this sentence is a lot more general, as the distributions of all the digits, and even that of the whole mantissa, can be derived. Newcomb did not write explicitly these formulae, although it is certain that he was fully aware of them, as he gave the following table of probability for the first two digits:

d	$\Pr(D_1 = d)$	$\Pr(D_2 = d)$
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

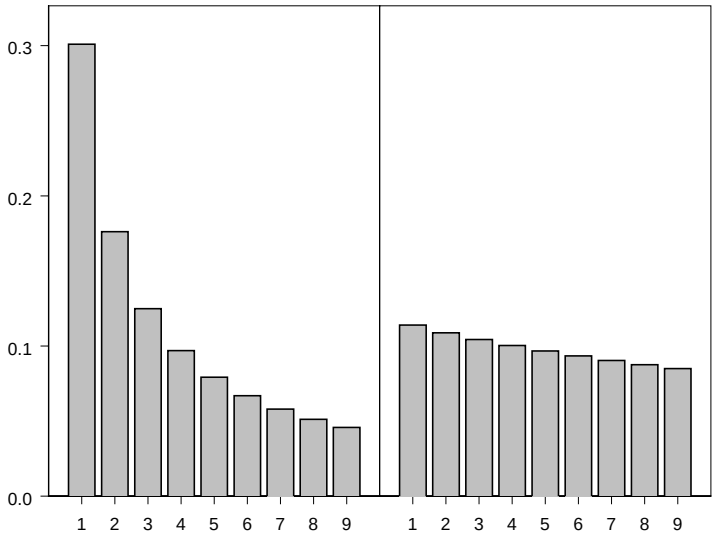


Figure 1.2: The distributions of the first and second digits as derived by Newcomb

It can thus be observed that the distribution of the second digit is less left skewed than that of the first digit, and it is also more linear. Newcomb pointed out this fact, and also that according to 'his' law the distribution of a digit converges to the uniform as its position increases in the mantissa, i.e.:

> In the case of the third figure the probability will be nearly the same for each digit, and for the fourth and following ones the difference will be inappreciable.

On the whole, Newcomb's article has a kind of genius sense, but as often in that kind of discovery it is not well explained at all. Moreover the discovery is totally counter-intuitive, so actually he would probably have been considered as a bit weird if his article had drawn any attention...

1.2 Benford's experiment

Newcomb's article was not well recognized at all, perhaps because of its apparent lack of mathematical background. However, about half a century later, a physicist at General Electric somehow 'rediscovered' the phenomenon, and so naturally the law took his name, unfortunately for his predecessor.

Like Newcomb before him, Frank Benford observed the difference of dirtiness on the logarithm pages. He then tried to reproduce in a now famous experiment what these tables were used for (see [Benford]). He collected some 20,229 observations of "natural" numbers, some coming from strict mathematical rules like sqare roots of integers, others from physics like constants and measurements, but also some from 'weird' datasets like the first 342 street addresses of *American Men of Science*, every number from an issue of *Readers Digest*, etc...

He computed the frequencies of the first digit for each different dataset he used, and also the average of all the datasets together. The results were summarized in the following table:

Title	1	2	3	4	5	6	7	8	9	Samples
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
Newspaper items	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
Blackbody	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
$n^1, n^2, \ldots, n!$	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
Probable Error ($\pm$)	0.8	0.4	0.4	0.3	0.2	0.2	0.2	0.2	0.3	

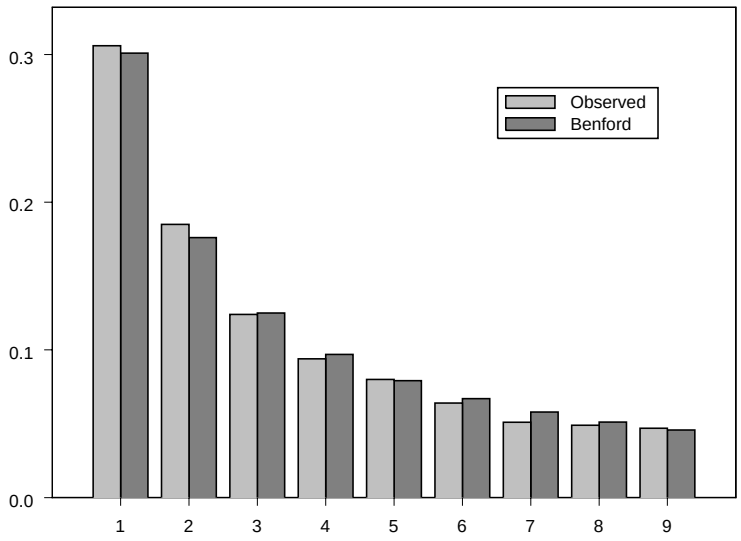


Figure 1.3: Average of all frequencies in Benford's experiment

Benford observed that the logarithmic law was better fitted by the more random numbers in his experiment ("those outlaw numbers without known relationships"), like the numbers taken from newspapers, the street adresses, and the physical measurements (air presssure, black body radiation, etc...), than the more deterministic ones, like the square roots of integers. However, what has to be pointed out is that it is the average which has the best fit to the logarithmic law.

1.3 Different attempts to explain Benford's law

Benford himself tried to explain the phenomenon by investigating the set of natural integers, in an attempt to prove that it comes naturally from our number system. As a start, he tried to prove that the set of integers that have one as first digit (i.e. $\{1, 11, 12, \ldots 100, 101, \ldots\}$) has a 'probability' of $\log 2$ among the integers. The problem that he encountered (and that many encountered after him) is that this set has no asymptotic natural frequency, i.e. the limit:

$$\lim \frac{1}{n} \# \{i \in \mathbb{N} \mid i \leq n \text{ and } d_1(i) = 1\}$$

does not exist.

If one tried to represent the behaviour of this sequence (extrapolating it between integers), this would look like this:

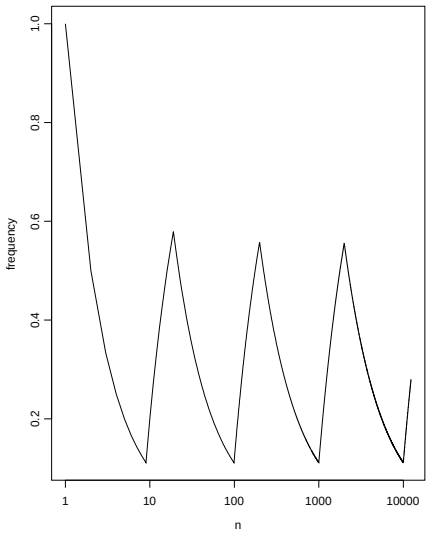


Figure 1.4: Frequency of appearance of the first digit 1 in the set $\{1, \dots n\}$

So the limit does not exist, the sequence is oscillating on a logarithmic scale between two extrema that are not even constant (even if it cannot be clearly seen on figure 1.4), the minima decreasing towards $\frac{1}{9} = 0.1111\dots$ and the maxima towards $\frac{5}{9} = 0.5555\dots$.

There are various ways to define a limit of such a sequence, and several of them lead to the desired $\log 2$. In fact many of the writers that followed adopted the same start as Benford, and different integration schemes were proven to be consistent with Benford's Law. Raimi has an excellent review of the 'old' litterature (see [Raimi]).

But the problem is then always the same for those explanations:

- Firstly, the integration methods are not unique, and of course plenty of them *do not* lead to Benford's law.
- Secondly, the approach itself is a bit dubious, since it aims at proving that the law is valid for the whole number system (so is completely universal), and of course lots of data sets -even 'natural' ones- *do not* confirm to Benford's law.

The first major breakthrough in that context was apparently made by Pinkham (who however attributes the idea to Hamming), in 1961 (see [Pinkham]). This very simple idea was: "if there's a universal law in nature then it should appear

whatever units are used to count". For example, if Benford's law is observed in a financial dataset expressed in dollars, then it should also appear in the same dataset expressed in francs. The action of converting from one unit to another is of course a scaling by a constant, and Pinkham not only discovered that the logarithmic law is invariant by scaling, but also that it is the only law that has such a property, which he naturally called scale-invariance (see 2.3.2 for the proofs).

There were still some enormous problems in that reasoning, though. The basic hypothesis was "if there's a universal law", and of course it is rather hard or even impossible to admit. And also, on a more mathematical ground, it is easy to see that *there is no scale invariant probability measure on the Borels*, since then for example the probability of $(0,1)$ would equal that of $(0,s)$, for every s, which is impossible if $(0,1)$ has a non zero probability. But the idea was there, and with the many datasets that were then known to follow Benford's Law, the path was set for different -and maybe more rigorous- approaches...

1.4 Recent developments

After the 1980s, a period during which few articles were published about the subject, the beginning of the 90s seems like a renewal for Benford's Law. This is perhaps due to the 'discovery' of a new important field of application, i.e. fraud detection (see below, section 1.6) . One of the first key papers to be written on the subject was Mark Nigrini's thesis (supported in 1992 at the Department of Accounting, University of Cincinatti).

In the 90s, one of the major contributions to the analysis of Benford's Law is certainly due to Theodore Hill, who set up a correct probability framework for Benford's Law, extended the idea of scale-invariance to base-invariance (why should a "universal law" be dependent of the base in which the numbers are written?), and introduced a new way of considering Benford's Law, as explained below.

Beforehand, many of the previous works on Benford's Law were somehow quoting "mystical" reasons when explaining the existence of the law. Actually many authors hypothesized a sort of behind-the-scene universal law, or which is the same, a universal table of constants, and tried to conjecture from this hypothesis.

However, as it was explained above in section 1.3, this cannot be proved at all. A more natural approach, and indeed so natural that surprisingly nobody explored it before Hill in 1995, is to think of data as being a mixture from *different distributions*. Actually this approach seems to be relevant to Benford's experiment, where the data came from more than 20 various distributions. Hill simply linked this idea with scale- and base-invariance to make a consistent -but unfinished- explanation of Benford's Law. Some of the later authors (see [Leemis *et al.*]) have tried to complete it, but there is still a lot of work to do in that domain. Section 2.3 tries to provide an insight of Hill's mathematical work on Benford's Law.

1.5 Empirical evidence

Distinct from the theoretical attempts at explaining the law, there had always been lots of publications about newly discovered Benford datasets. In fact not surprisingly many of them were made by physicists or computer scientists. Considering the number of datasets in the universe compared with the number of people aware of Benford's Law, there still must be a lot to find... However, these datasets were often found by luck, so the chances of someone looking on purpose for a new kind of Benford dataset succeeding are not necessarily high...

The following list, which does not include the ones found by Benford, does not pretend to be complete, but tries to offer a view as broad as possible:

- **Physical constants**: The list of the most used constants in physics has been found by several authors to offer a rather good fit to Benford's Law (see [Knuth] and [Burke & Kincanon]). This is not a really convincing example (in the author's opinion), since in general there are too few of them to draw a strong statistical conclusion.

- **Physical measurements**: Some physicists and engineers remarked that the first digits of their own data had the logarithmic distribution. In engineering for instance Becker found that the failure rates and the MTTFs (Mean Time To Failure), taken from classical tables, often satisfy Benford's Law (see [Becker]). In atomic physics, Buck, Merchant and Perez observed that the first digits of both the predicted and the observed values of 477 radioactive half-lives have the logarithmic distribution (see [Buck *et al.*]).

- **Scientific calculations**: In case of long series of floating-point operations, the mantissae are known to follow a logarithmic distribution. Actually some specialists just wondered why it is not considered more often, given the very high frequency of appearance of the phenomenon (see [Hamming] and [Knuth]).

- **Financial and accounting data**: Many Benford datasets have been found in this domain. In accounting, Benford's Law appears very often, as "lists of items, such as accounts receivable or payable, inventory counts, fixed asset acquisitions, daily sales, and disbursements, should follow Benford's Law" (see [Nigrini & Mittermaier]). In finance, Ley for example found that "the series of 1-day returns on the Dow-Jones Industrial Average Index (DJIA) and the Standard and Poor's Index (S&P) reasonably agrees with Benford's Law" (see [Ley]).

- **Populations**: Census data is another important type of Benford datasets. In an apparently unpublished study Nigrini and Wood found that the populations of the 3141 U.S. counties in 1990 were a good fit to Benford's Law (see [Hilla]). In the present report some other demographic data are considered (in chapter 3).

1.6 Applications

At the beginning not many applications were seen by scientists. Benford's Law was seen as a kind of mathematical nicety without any practical returns. Then in the 70s and with the arrival of computers some applications were imagined, and lately in the 90s a new important area of application was unveiled, as described below. Hence now three major domains, each based on a different kind of dataset, can be distinguished:

- **Computer design**: When designing a computer or writing routines, distributions of operands have to be considered. Coding real numbers requires for example to decide with which number of bits the mantissa will be described (this is linked with confidence intervals). A certain pattern in the mantissa will lead to different choices, in order to optimize storage and/or processing speed. Schatte, Knuth, and other various authors in the 70s and 80s have thus investigated what they called the "logarithmic computer" (see [Schatte], [Knuth] and [Hilla], and the papers referenced there), but actually with the very fast improvement of technology that occurs in computing equipment this seems a bit 'old-fashioned'...

- **Modelling**: This application was apparently imagined by Varian in 1972 (see [Varian]). It is based on the following very simple idea: if a certain set of values follows Benford's Law, then models for the corresponding predicted values should also follow Benford's Law. Hill qualified this type of application as "Benford-in-Benford-out" ([Hilla]). This includes (for example) models for demographic growth, financial indexes, or any other Benford dataset...

- **Fraud detection**: This is a very trendy application, linked with such topics as data mining, expert systems, neural networks... Basically, it comes from the observation that manipulated -or fraudulent- data do not tend to confirm to Benford's Law, whereas unmanipulated data do. Benford's Law also helps to spot duplicate entries in databases, like double payments, and is thus interesting for most companies. The first use of Benford's Law in that area was conducted by the District Attorney's office in Brooklyn, New York, that charged (sucessfully) seven companies for fraud using a computer programme made by Nigrini and implementing Benford goodness-of-fit tests (this was announced in the *Wall Street Journal* on July, 10th 1995). From then on use of Benford's Law has been widespread in the accounting world (4 of the "Big Five", the most famous U.S. accounting firms, use it), some computer packages have been designed on it, and it has lead to a kind of fraud detection system called "Digital Analysis", that relies not only on Benford's Law, but also on other tests of digits. Some of the interesting papers on the subject are [Nigrini & Mittermaier] and [Nigrini].

Chapter 2

Main mathematical results

2.1 The significant-digit law

The aim of this section is to define formally Benford's law and to give some basic results on the involved distributions.

2.1.1 The first digit distribution

Definition 2.1.1 *The logarithmic (discrete) density function for the first digit D is defined by:*

$$\boxed{\Pr(\mathbf{D} = \mathbf{d}) = \log\left(\mathbf{1} + \frac{\mathbf{1}}{\mathbf{d}}\right)}$$

where $d \in \{1, \ldots, 9\}$.

The corresponding distribution function follows easily:

$$\begin{aligned}
\Pr(D \leq d) &= \sum_{1 \leq d' \leq d} \Pr(D = d') = \sum_{1 \leq d' \leq d} \log\left(1 + \frac{1}{d'}\right) \\
&= \log\left(\prod_{1 \leq d' \leq d} \left(1 + \frac{1}{d'}\right)\right) \\
&= \log\left(\left(1 + \frac{1}{1}\right)\left(1 + \frac{1}{2}\right) \cdots \left(1 + \frac{1}{d}\right)\right) \\
&= \log\left(\frac{2}{1} \times \frac{3}{2} \times \cdots \times \frac{d+1}{d}\right) \\
&= \log(d+1) \qquad (1)
\end{aligned}$$

where $d \in \{1, \ldots, 9\}$.

Remark 2.1.2 $\Pr(D \leq 9) = 1$, *so the probability density function defined in definition 2.1.1 is well defined on* $\{1, \ldots, 9\}$.

2.1.2 The joint distribution of digits and the k^{th} digit distribution

Definition 2.1.3 *The logarithmic joint distribution of the first significant digits* $D_1, D_2, \ldots, D_k$ *(for every* $k \in \mathbb{N}^*$*) is defined by:*

$$\Pr(\mathbf{D_1 = d_1}, \ldots, \mathbf{D_k = d_k}) = \log\left(\mathbf{1} + \left(\sum_{\mathbf{i=1}}^{\mathbf{k}} \mathbf{10^{k-i} d_i}\right)^{\mathbf{-1}}\right)$$

where $d_1 \in \{1, \ldots, 9\}$, *and all other* $d_j \in \{0, \ldots, 9\}$.

Example 2.1.4 *The probability of any combination of digits can be very easily found using this expression, e.g.*

$$\Pr(D_1 = 1, D_2 = 2, D_3 = 3) = \log\left(1 + \frac{1}{123}\right) = \log\left(\frac{124}{123}\right)$$

The discrete density function for the k^{th} digit D_k can be derived using definition 2.1.3:

$$\begin{aligned} \Pr(D_k = d_k) &= \sum_{\substack{1 \leq d_1 \leq 9 \\ 0 \leq d_2 \leq 9 \\ \cdots \\ 0 \leq d_{k-1} \leq 9}} \Pr(D_1 = d_1, D_2 = d_2 \ldots, D_k = d_k) \\ &= \sum_{\substack{1 \leq d_1 \leq 9 \\ 0 \leq d_2 \leq 9 \\ \cdots \\ 0 \leq d_{k-1} \leq 9}} \log\left(1 + \left(\sum_{i=1}^{k} 10^{k-i} d_i\right)^{-1}\right) \end{aligned} \tag{2}$$

Remark 2.1.5 *If the significant digits follow logarithmic laws, they are not independent! For example,* $P(D_1 = 1, D_2 = 2) = \log\left(\frac{13}{12}\right) = 0.035$, *whereas* $P(D_1 = 1) \times P(D_2 = 2) = \log(2) \times \left(\log\left(\frac{13}{12}\right) + \log\left(\frac{23}{22}\right) + \cdots + \log\left(\frac{93}{92}\right)\right) = 0.033$.

2.1.3 The mantissa distribution

Lemma 2.1.6 *The logarithmic density in definition 2.1.3 can be generalized in a continuous way for the mantissa* M *in the following form:*

$$\Pr(\mathbf{M \leq m}) = \mathbf{\log(m)}$$

where $m \in [1, 10)$.

Proof.

- First consider the case where $m = d_1$, i.e. m has only one significant digit:

$$\Pr(M \leq m) = \begin{cases} 0 \text{ if } d_1 \leq 1 \text{ (} \Pr(M = 1) = 0 \textit{ in a continuous case } \text{)} \\ \\ \Pr(D_1 \leq d_1 - 1) = \log(d_1) = \log(m) \text{ if } 1 < d_1 \leq 10 \end{cases}$$

- Then suppose $m = d_1 d_2 \ldots d_k$ in decimal notation, i.e. $m = \sum_{i=1}^{k} 10^{-(i-1)} d_i$, with $d_1 > 1, d_2 > 0, \ldots, d_k > 0$:

$$\begin{aligned} \Pr(M \leq m) &= \Pr(D_1 \leq d_1 - 1) \\ &\quad + \Pr(D_1 = d_1, D_2 \leq d_2 - 1) \\ &\quad + \cdots \\ &\quad + \Pr(D_1 = d_1, D_2 = d_2, \ldots, D_{k-1} = d_{k-1}, D_k \leq d_k - 1) \end{aligned} \tag{3}$$

$$\begin{aligned} &= \Pr(D_1 \leq d_1 - 1) \\ &\quad + \sum_{0 \leq d_2' \leq d_2 - 1} \Pr(D_1 = d_1, D_2 = d_2') \\ &\quad + \cdots \\ &\quad + \sum_{0 \leq d_k' \leq d_k - 1} \Pr(D_1 = d_1, D_2 = d_2 - 1, \ldots, D_k = d_k') \end{aligned}$$

$$\begin{aligned} &= \log(d_1) \\ &\quad + \sum_{0 \leq d_2' \leq d_2 - 1} \log\left(1 + \frac{1}{10 d_1 + d_2'}\right) \\ &\quad + \cdots \\ &\quad + \sum_{0 \leq d_k' \leq d_k - 1} \log\left(1 + \left(\sum_{i=1}^{k-1} 10^{k-i} d_i + d_k'\right)^{-1}\right) \end{aligned}$$

By analogy with the derivation of the first digit distribution (1), we find:

$$\begin{aligned} \Pr(M \leq m) &= \log(d_1) + \log\left(\frac{10 d_1 + d_2}{10 d_1}\right) + \cdots + \log\left(\frac{\sum_{i=1}^{k} 10^{k-i} d_i}{\sum_{i=1}^{k-1} 10^{k-i} d_i}\right) \\ &= \log\left(\frac{\sum_{i=1}^{k} 10^{k-i} d_i}{10^{k-1}}\right) \\ &= \log\left(\sum_{i=1}^{k} 10^{-(i-1)} d_i\right) \\ &= \log(m) \end{aligned}$$

- Now if one (or several) of the d_j ($j > 1$) are null or if $d_1 = 1$, the result above is still true.

  Say $m = d_1 d_2 \dots d_{j-1}\, 0\, d_{j+1} \dots d_k$ for example. Then there is no j^{th} term in the sum (3), and this leads to the following expression:

$$\begin{aligned}
\Pr(M \leq m) &= \log(d_1) + \cdots \\
&\quad + \log\left(\frac{\sum_{i=1}^{j-1} 10^{j-1-i} d_i}{\sum_{i=1}^{j-2} 10^{j-1-i} d_i}\right) + 0 + \log\left(\frac{\sum_{i=1}^{j+1} 10^{j+1-i} d_i}{\sum_{i=1}^{j} 10^{j+1-i} d_i}\right) + \cdots \\
&\quad + \log\left(\frac{\sum_{i=1}^{k} 10^{k-i} d_i}{\sum_{i=1}^{k-1} 10^{k-i} d_i}\right) \\
&= \log\left(\frac{1}{10^{j-2}} \times \frac{\sum_{i=1}^{j-1} 10^{j-1-i} d_i}{\sum_{i=1}^{j} 10^{j+1-i} d_i} \times \frac{1}{10^{k-j-1}} \times \sum_{i=1}^{k} 10^{k-i} d_i\right) \\
&= \log\left(\frac{1}{10^{j-2}} \times \frac{1}{10^2} \times \frac{1}{10^{k-j-1}} \times \sum_{i=1}^{k} 10^{k-i} d_i\right) \\
&= \log\left(\frac{1}{10^{k-1}} \times \sum_{i=1}^{k} 10^{k-i} d_i\right) \\
&= \log(m)
\end{aligned}$$

  This can be easily extended to the case where several d_j are null or if $d_1 = 1$.

■

Remark 2.1.7 *In fact it is easier and more instructive to think the other way round (i.e. to derive the distribution of the k^{th} digit from the distribution of the mantissa):*

$$\begin{aligned}
\Pr(D_1 = d_1, D_2 = d_2, \dots, D_k = d_k) &= \Pr(d_1 + d_2 10^{-1} + \dots + d_k 10^{-k+1} \leq M < \\
&\quad d_1 + d_2 10^{-1} + \dots + (d_k + 1) 10^{-k+1}) \\
&= \log(d_1 + d_2 10^{-1} + \dots + (d_k + 1) 10^{-k+1}) \\
&\quad - \log(d_1 + d_2 10^{-1} + \dots + d_k 10^{-k+1}) \\
&= \log\left(\frac{d_1 + d_2 10^{-1} + \dots + (d_k + 1) 10^{-k+1}}{d_1 + d_2 10^{-1} + \dots + d_k 10^{-k+1}}\right) \\
&= \log\left(1 + \left(\sum_{i=1}^{k} 10^{k-i} d_i\right)^{-1}\right)
\end{aligned}$$

This way of thinking is very general and can be used in a lot of situations, like in example 2.1.4.

The following key definition of Benford's law can now be stated:

Definition 2.1.8 *A random variable X satisfies Benford's law for the mantissa if $M = M(X)$ follows the logarithmic mantissa distribution.*
A random variable X satisfies Benford's law for the k^{th} digit if $D_k = D_k(X)$ follows the logarithmic k^{th} digit distribution.

2.1.4 Variate generation

It is very easy to generate random variables that follow Benford's distributions.

Recall that the cumulative distribution for the mantissa M is:

$$\Pr(M \leq m) = \log(m)$$

for $m \in [1, 10)$.

So, by a straightforward inversion, a random variable M that follows Benford's mantissa distribution can be generated by:

$$\mathbf{M \leftarrow 10^U} \tag{4}$$

where $U \sim \mathcal{U}(0, 1)$.

As a consequence, a random variable D_1 that follows Benford's first digit distribution can be generated by:

$$D_1 \leftarrow \lfloor 10^U \rfloor$$

The generation method (4) helps to explain Newcomb's statement, which is that **the mantissae of the logarithms of numbers (which satisfy Benford's law) are uniformly distributed** (see section 1.1):

Let X be the random variable that represents the number whose logarithm is searched for, M its mantissa, and S the integer random variable such that $10^S \leq X < 10^{S+1}$ (S is the scale of X in base 10, i.e $X = M \times 10^S$).
Hence: $M = X \times 10^{-S} = 10^{\log X - S}$
So, according to the generation procedure (4), if $\log X - S \sim \mathcal{U}(0, 1)$, then M follows Benford's mantissa distribution (so X satisfies Benford's law), and conversely.

Remark 2.1.9 *In fact, $\log X - S$ represents the result of the logarithm table, i.e. the floating part of the logarithm (or equivalently the logarithm modulo 1), so it is not exactly the mantissa of the logarithm as Newcomb stated (it was clear in his mind, though, only the sentence is a bit ambiguous).*

Remark 2.1.10 *This suggests a systematic method to check whether a random variable X satisfies Benford's law or not:*
X satisfies Benford's law if and only if $\log X - \lfloor \log X \rfloor \sim \mathcal{U}(0, 1)$ or equivalently if $\log X \mod 1 \sim \mathcal{U}(0, 1)$.

2.1.5 Some distributions that exactly satisfy Benford's law

A distribution is said to satisfy Benford's law if its corresponding random variable does. It is relatively easy to find such distributions; in fact many of them seem to satisfy it.

For example, suppose we are looking for a distribution on $[1, 10)$ satisfying Benford's law for the first digit, say f_X. Then for $d \in \{1, \ldots, 9\}$ we have:

$\Pr(D_1 = d) = \Pr(d \leq X < d+1) = \int_d^{d+1} f_X(x)dx = \log\left(\frac{d+1}{d}\right)$

So a natural way of constructing f_X is to choose $f_X(x) = \frac{1}{x \ln 10}$. In fact this distribution can be found using (4), and it satisfies Benford's law for the whole mantissa, as it is that of a random variable $X = 10^U$ where $U \sim \mathcal{U}(0,1)$ (the corresponding cumulative distribution function F_X is: $F_X(x) = \int_1^x \frac{1}{t \ln 10} dt = \log x$, and inverting it yields $X = 10^U$)

This rather simple distribution is only useful to give a clue about how to find other distributions that satisfy Benford's law. If a distribution on $[10^a, 10^b)$ is searched for, it is thus natural to think of that of 10^U, where $U \sim \mathcal{U}(a, b)$, which is $f_X(x) = \frac{1}{(b-a)x \ln 10}$.

And indeed we find:

$$\begin{aligned}
\Pr(D_1 = d) &= \Pr(10^a \times d \leq X < 10^a \times (d+1)) \\
&\quad + \Pr(10^{a+1} \times d \leq X < 10^{a+1} \times (d+1)) \\
&\quad + \cdots \\
&\quad + \Pr(10^{b-1} \times d \leq X < 10^{b-1} \times (d+1)) \\
&= \frac{1}{b-a} \times \left(\log\left(\frac{10^a \times (d+1)}{10^a \times d}\right) \right. \\
&\quad + \log\left(\frac{10^{a+1} \times (d+1)}{10^{a+1} \times d}\right) \\
&\quad + \cdots \\
&\quad \left. + \log\left(\frac{10^{b-1} \times (d+1)}{10^{b-1} \times d}\right)\right) \\
&= \frac{1}{b-a} \times \left((b-a) \times \log\left(\frac{d+1}{d}\right)\right) \\
&= \log\left(\frac{d+1}{d}\right)
\end{aligned}$$

Using the same idea, several other distributions were found satisfying Benford's law. These distributions are of the form 10^W, with W a random variable whose support is an interval between integers. Among these we can quote (see [Leemis *et al.*]):

1. $f_W(w) = \begin{cases} w & \text{if} \quad 0 \leq w \leq 1 \\ 2 - w & \text{if} \quad 1 \leq w \leq 2 \end{cases}$

   (a triangular distribution on $[0, 2]$)

It can now be checked that Benford's law is satisfied by deriving the cumulative distribution function of $Z = W - S$, where $S = \lfloor W \rfloor$ (c.f. remark 2.1.10). Conditioning by S gives, for all $z \in [0,1]$:

$$\begin{aligned} F_Z(z) &= \Pr(S=0) \times \Pr(W \leq z \mid S=0) \\ &\quad + \Pr(S=1) \times \Pr(W-1 \leq z \mid S=1) \\ &= \Pr(W \leq z \ \cap \ 0 \leq W \leq 1) \\ &\quad + \Pr(W \leq z+1 \ \cap \ 1 \leq W \leq 2) \\ &= \Pr(W \leq z) + \Pr(1 \leq W \leq z+1) \\ &= \int_0^z w \, dw + \int_1^{z+1} (2-w) \, dw \\ &= \frac{z^2}{2} + \left(z - \frac{z^2}{2} \right) \\ &= z \end{aligned}$$

Hence $Z \sim \mathcal{U}[0,1]$, and according to remark 2.1.10, $X = 10^W$ satisfies Benford's law.

2. $f_W(w) = \begin{cases} 1 - w^2 & \text{if} \quad -1 \leq w \leq 0 \\ (w-1)^2 & \text{if} \quad 0 \leq w \leq 1 \end{cases}$

(a non-symmetric distribution on $[-1,1]$)

It can be shown that $X = 10^W$ satisfies Benford's law using the same process as above:

$$\begin{aligned} F_Z(z) &= \Pr(S=-1) \times \Pr(W+1 \leq z \mid -1 \leq W \leq 0) \\ &\quad + \Pr(S=0) \times \Pr(W \leq z \mid 0 \leq W \leq 1) \\ &= \Pr(W \leq z-1) + \Pr(0 \leq W \leq z) \\ &= \int_{-1}^{z-1} (1-w^2) \, dw + \int_0^z (w-1)^2 \, dw \\ &= \left(-\frac{z^3}{3} + z^2 \right) + \left(\frac{z^3}{3} - z^2 + z \right) \\ &= z \end{aligned}$$

2.2 Convergence to the uniform distribution

The reason why Benford's law is only useful for the first digits (often one or two in practice) is that the distribution of the k^{th} digit tends to the uniform distribution on $\{0, \ldots, 9\}$ exponentially fast as k increases. Newcomb had already remarked the phenomenon.

In fact, this property was proven through direct calculation by Diaconis (see [Diaconis]), but there is a more general result that is due to Hill and that has apparently not been published yet.

Conjecture 2.2.1 *Let X be* **any** *continuous random variable with bounded piecewise smooth density function. Then the distribution of $D_k(X)$ approaches the uniform distribution as k goes to infinity.*

Proof.

There is currently no formal proof for that proposition, but it is rather easy to sketch an informal one.

Without loss of generality, suppose the base used for describing numbers is 2.

Take an interval, say $[1, 2]$, and draw an arbitrary bounded continuous function, say an exponential, on that interval. This function can then always be normalized to define a probability density function.

In base 2, $[1, 2]$ is written $[1.00, 10.00]$. Among all the real numbers in $[1, 2]$, those which have 0 as their second (binary) digit are between 1.00 and 1.10 (i.e 1 and 1.5 in decimal notation). Those which have 0 as their third (binary digit) are between 1.00 and 1.01 (1 and 1.25) and between 1.10 and 1.11 (1.5 and 1.75), etc...

As the process goes to the infinite, it will split in half the complete integral, and hence the probability of observing a 0 will be the same as observing a 1 (uniform distribution).

The set of figures 2.2 describes the process in image, with a simple (normalized) exponential function between 1 and 2.

For any interval the behaviour will be esentially the same. $[1, 2]$ in base 2 was chosen for simplicity, but it will work for any interval, asymptotically the interval will be divided in half.

For the decimal digits the behaviour will also be the same, except there will be 10 subdvisions of the interval instead of 2.

Actually the hypotheses are certainly a bit too strong, as it should work for any "non-pathological" function (i.e. any usual probability density function).

The formal proof should probably use some results about fractals, or may even have been published in that domain.

■

Remark 2.2.2 *This result is useful because it tells that the farther the mantissa is looked at, the less information it provides. This is in a sense very intuitive, because if one had potentially to look to the infinity in the significant digits to find a piece of information, then there would be no reliable probability calculus...*

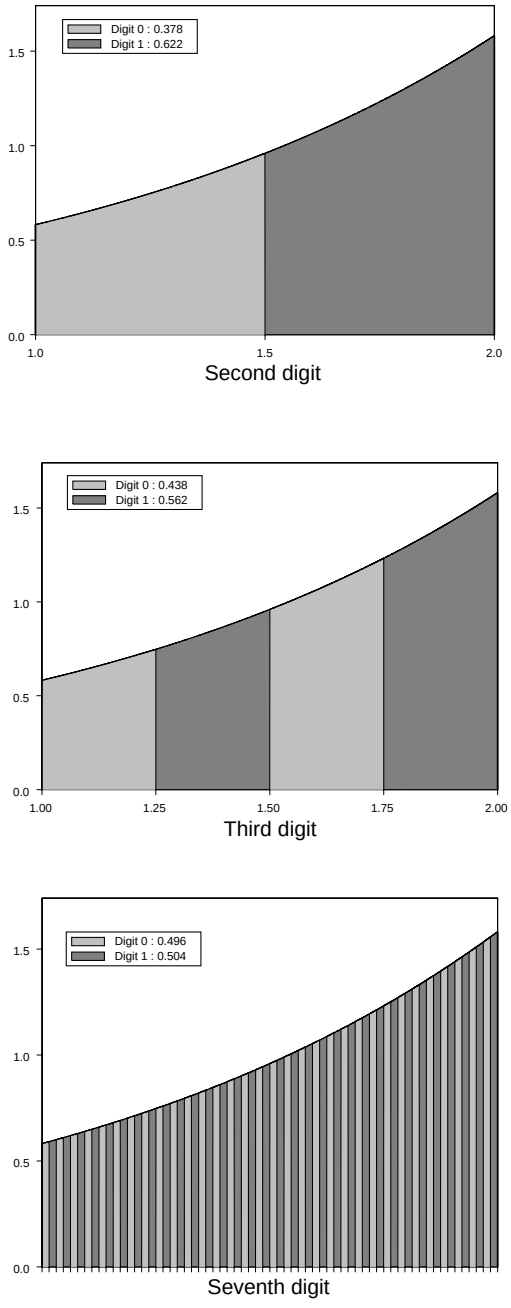


Figure 2.1: Convergence to the uniform distribution

And for the still skeptical ones, here are the computed probabilities that the k^{th} (binary) digit equals zero for 8 different distributions conditioned between 1 and 2:

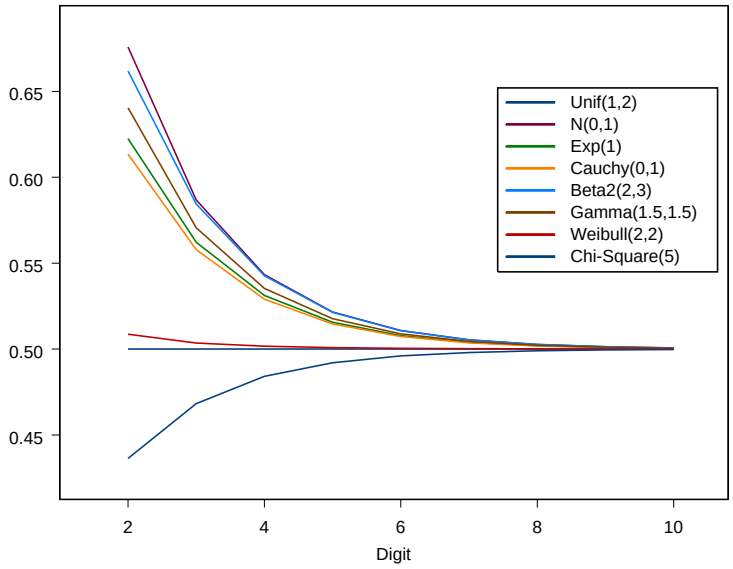


Figure 2.2: Distribution of binary digits of 8 distributions defined on $[1, 2]$

2.3 A mathematical explanation of Benford's law: Hill's theorems

This section tries to give some insights on Hill's work (see [Hilla], and also [Hillb], [Hillc]), which seems to be currently one of the most convincing explanations of the recurrence of Benford's law in nature phenomenons.

2.3.1 The proper probability space

Defining a proper probability space is one of the keys to understand and to study Benford's law. One of the problems that quickly appears when constructing such a space is that the event space is not a 'normal' one, i.e. not the set of Borels of $\mathbb{R}$. For example, a single interval never contains all the real numbers whose mantissae belong to a certain Borel set (think of $[10, 20)$: every number in $[10, 20)$ has its mantissa in $[1, 2)$, but of course every number in

$[1,2), [100,200), [1000,2000)...$ has also its mantissa in $[1,2)$). In other words, a probability cannot be assigned to a single interval if one only looks at the corresponding set of mantissae.

However, following that example, it is rather easy to see that the set of real numbers $\bigcup_{n=-\infty}^{\infty} B \times 10^n$ contains all the positive numbers whose mantissae belong to B, where B is a Borel of $[1,10)$. An observer who would only look at mantissae could assign a probability to this set. So a natural definition of the measurable space associated with a Benford-like experiment, where only positive real numbers are sampled, is $(\mathbb{R}^+, \mathcal{M})$ where the event space $\mathcal{M}$ is defined in the following way:

Definition 2.3.1 *The event space $\mathcal{M}$, called the* **mantissa algebra**, *is defined by:*

$$\boxed{\mathcal{M} = \left\{ \bigcup_{\mathbf{n}=-\infty}^{\infty} \mathbf{B} \times \mathbf{10^n} \textbf{ \textit{for all Borel} } \mathbf{B} \subseteq [\mathbf{1}, \mathbf{10}) \right\}}$$

The mantissa algebra $\mathcal{M}$ is of course a σ-algebra, as a sub-σ-field of the Borels. The following lemma is useful to understand the structure of the mantissa algebra.

Lemma 2.3.2 *The main properties of the mantissa algebra are the following:*

1. *Every non-empty set in $\mathcal{M}$ is infinite with accumulation points at 0 and at $+\infty$ (i.e. in any set S in $\mathcal{M}$ it is always possible to find arbitrary large and arbitrary small non-zero numbers)*
2. *$\mathcal{M}$ is closed under scalar multiplication (for all $s > 0, S \in \mathcal{M} \Rightarrow sS \in \mathcal{M}$),*
3. *$\mathcal{M}$ is closed under integral roots, but not integral powers (for all $m \in \mathbb{N}$, $S \in \mathcal{M} \Rightarrow S^{1/m} \in \mathcal{M}$),*
4. *$\mathcal{M}$ is self-similar in the sense that for all $m \in \mathbb{N}$, $S \in \mathcal{M} \Rightarrow 10^m S \in \mathcal{M}$.*

The properties 1, 2 and 4 are easy to understand, whereas property 3 needs more attention. Here is the proof of property 3:

Proof.

First, proove that $S^{\frac{1}{m}} \in \mathcal{M}$.

Let $s \in S^{\frac{1}{m}}$. Then by definition 2.3.1:

$$\exists\, n \in \mathbb{N} \text{ and a Borel } B \subseteq [1,10) \text{ s.t. } s^m \in B \times 10^n$$

So by the definition of a Borel set, $\exists a, b \in [1,10)$ s.t. $a \times 10^n \leq s^m < b \times 10^n$. Hence by the Euclidean division of n by m:

$$\exists\, q \in \mathbb{N},\ k \in \{0, \ldots, m-1\} \text{ s.t. } a^{\frac{1}{m}} \times 10^{\frac{k}{m}+q} \leq s < b^{\frac{1}{m}} \times 10^{\frac{k}{m}+q}$$

Let

$$B_m = \bigcup_{k=0}^{m-1} \left[a^{\frac{1}{m}}, b^{\frac{1}{m}} \right) \times 10^{\frac{k}{m}} \tag{5}$$

B_m is clearly a Borel set of $[1, 10)$, and $s \in \bigcup_{q=-\infty}^{\infty} B_m \times 10^q$, so $S^{\frac{1}{m}} \in \mathcal{M}$.

Now, consider the set S_1 of all real numbers whose first digit is a one:

$$S_1 = \bigcup_{q=-\infty}^{\infty} [1, 2)10^q$$

The set S_1^2 is the following:

$$S_1^2 = \bigcup_{q=-\infty}^{\infty} [1, 4)10^{2q}$$

Thus S_1^2 does not belong to $\mathcal{M}$, since for example it includes $[1, 4)$ but not $[10, 40)$.

■

Now that the probability space is defined and that its properties are known, the next step is to state some reasonable hypotheses and show that Benford's law comes from those hypotheses. Property 2 leads to the hypothesis of scale invariance, whereas property 3 is a key to the weaker hypothesis of base invariance, as it will be seen in next subsection.

2.3.2 Scale invariance

Scale invariance is one of the simplest hypothesis that comes when thinking of the "universality" of Benford's law. Suppose that there is a law that would somehow appear in "natural" data sets whatever system of number is chosen. The first thing that comes to mind is that it would be expected to be independent of the unit system, i.e. rescaling all the numbers by a constant should not affect the probability measure. This property is called scale invariance, and is formally defined in the following way (recall that $\mathcal{M}$ is closed under scalar multiplication, lemma 2.3.2, property 2):

Definition 2.3.3 *A probability measure P on the mantissa algebra $\mathcal{M}$ is scale invariant if*

$$\forall\, s \in \mathbb{R}, \forall\, S \in \mathcal{M}, P(S) = P(sS)$$

The following theorem shows that scale invariance is a characterization of Benford's law, and thus makes it very peculiar.

Theorem 2.3.4 *A probability measure P on $(\mathbb{R}^+, \mathcal{M})$ is scale invariant if and if only P satisfies Benford's law.*

Proof.

Let P be a probability measure on $(\mathbb{R}^+, \mathcal{M})$, and $S_a = \bigcup_{n=-\infty}^{\infty} [1, 10^a) \times 10^n$ for an arbitary $a \in [0, 1)$ (a probability measure on $\mathcal{M}$ is entirely defined by its values on such sets).

Let $\overline{P}$ and $\hat{P}$ be the probability measures defined respectively on the measurable spaces $\left([0,1),\ \mathbb{B}[0,1\,]\right)$ and $\left([1,10),\ \mathbb{B}[1,10\,]\right)$ by:

$$\forall\, a \in [0,1),\ \overline{P}\,[0,a) = \hat{P}\,[1,10^a) = P(S_a) \tag{6}$$

This relation defines actually a useful 1:1 correspondence between the respective measurable spaces of $P, \hat{P}$, and $\overline{P}$.

Now for the proof itself. In the measurable space $(\mathbb{R}^+, \mathcal{M})$, P satisfies Benford's law if and if only $P(S_a) = a$ for all $a \in [0,1)$ (so according to (6) if only $\overline{P}$ is the uniform distribution on $[0,1]$, and if only $\hat{P}\,[1,10^a) = a$, $\forall\, a \in [0,1)$).

Now suppose P satisfies Benford's law, and prove that for all $s \in \mathbb{R}$:

$$\begin{aligned} P(sS_a) &= P\left(\bigcup_{n=-\infty}^{\infty} [s, s\times 10^a)\times 10^n\right) \\ &= P(S_a) \end{aligned}$$

Without loss of generality, s can be restrained to $[1,10[$ (else take s mod 10).

Two cases are now to be distinguished:

- If $s \times 10^a \leq 10$:

$$\begin{aligned} P(sS_a) &= \hat{P}\,[s, s\times 10^a) \\ &= \hat{P}\,[1, s\times 10^a) - \hat{P}\,[1,s) \\ &= \log(s\times 10^a) - \log(s) \\ &= a \\ &= P(S_a) \end{aligned}$$

- If $s \times 10^a > 10$: (since $s \leq 10$, $s\times 10^a$ is in $[10,100)$)

$$\begin{aligned} P(sS_a) &= \hat{P}\,[s,10) + \hat{P}\,([10, s\times 10^a) \mod 10\,) \\ &= (1-\log(s)) + \hat{P}\left[1, \frac{s\times 10^a}{10}\right) \\ &= 1 - \log(s) + \log\left(\frac{s\times 10^a}{10}\right) \\ &= 1 - \log(s) + \log(s) + a - 1 \\ &= a \\ &= P(S_a) \end{aligned}$$

Conversely, suppose P is scale invariant, i.e. $P(S_a) = P(sS_a)$ for all $s \in \mathbb{R}$ and $a \in [0,1)$, and show that P satisfies Benford's law, i.e. $P(S_a) = a$.

Let α be an arbitary irrational in $\mathbb{R}$. Then $P(S_a) = P(10^\alpha S_a)$, for all a in $[0,1)$. Without loss of generality, it can be supposed that $10^\alpha \in [1,10)$ (else take $10^\alpha \mod 10$).

The isomorphism defined in (6) then implies that:

$$\hat{P}[1,10^a) = \hat{P}\left([10^\alpha, 10^{a+\alpha}) \mod 10\right), \forall\, a \in [0,1)$$

[*Here the notation* $[a,b) \bmod 10$ *means* $[a,b)$ *if* $b \leq 10$, *or* $[a,10) \cup [1, b/10)$ *if* $b \in [10,100)$]

And, as a consequence:

$$\overline{P}[0,a) = \overline{P}\left([\alpha, a+\alpha) \mod 1\right), \forall\, a \in [0,1)$$

[*With the equivalent notation.*]

This last equality means that $\overline{P}$ is invariant by an irrational rotation on the unit circle. It has long been known that the unique distribution on $[0,1)$ that has such a property is the uniform (see for example [Weyl]), and consequently P satisfies Benford's law.

■

Remark 2.3.5 *This proof shows that the definition of scale invariance presented here in definition 2.3.3 is a bit too strong. In fact the strong scale invariance hypothesis can be reduced to the hypothesis of scale invariance by an arbitary number that is not a rational power of the base (here* α *is irrational, and the proof only uses scale invariance by* 10^α*).*

2.3.3 Base invariance

Base invariance is a more subtle hypothesis that also leads to Benford's law. If a certain law should appear in observing some "natural" datasets in base 10, then the idea is that this law should also appear if another base was used. Here the definition of the mantissa algebra $\mathcal{M}$ that was used before is going to be generalized to other bases, and the notation $\mathcal{M}_b$ will denote the mantissa algebra in an arbitrary base b (thus $\mathcal{M}_{10} = \mathcal{M}$). **All the properties, definitions, and theorems are essentially the same**, except that b replaces 10 (for example $\log_b$ replaces log in the probability distributions).

Now the following definition of base invariance can be presented (recall that $\mathcal{M}_b$ is closed under integral roots, lemma 2.3.2, property 3):

Definition 2.3.6 *A probability measure* P *on* $(\mathbb{R}^+, \mathcal{M}_b)$ *is base invariant if*

$$\forall\, m \in \mathbb{N}, \forall\, S \in \mathcal{M}_b, P(S) = P(S^{\frac{1}{m}})$$

Understanding why this definition is relevant to base invariance is not very easy at first sight. To motivate it, consider:

$$S = \bigcup_{q=-\infty}^{\infty} [\, b^x, b^y\,) \times b^q$$

where $x, y \in [0,1)$ (any left-closed interval of $[1,b)$ can be represented in the form $[b^x, b^y)$)

Use expression (5) in the proof of property 3 in lemma 2.3.2 (replace 10 by b):

$$\begin{aligned} S^{\frac{1}{m}} &= \bigcup_{q=-\infty}^{\infty} B_m \times b^q \\ &= \bigcup_{q=-\infty}^{\infty} \left(\bigcup_{k=0}^{m-1} [\, b^{\frac{x}{m}}, b^{\frac{y}{m}}\,)\, b^{\frac{k}{m}} \right) \times b^q \end{aligned}$$

Now replace b in the expression above by b^m:

$$\begin{aligned} S^{\frac{1}{m}} \quad \text{"} = \text{"} \quad & \bigcup_{q=-\infty}^{\infty} \left(\bigcup_{k=0}^{m-1} [\, b^x, b^y)\, b^k \right) \times b^{mq} \\ \text{"} = \text{"} \quad & \bigcup_{q=-\infty}^{\infty} \left(\bigcup_{k=0}^{m-1} [\, b^x, b^y)\, b^{mq+k} \right) \\ \text{"} = \text{"} \quad & \bigcup_{n=-\infty}^{\infty} [\, b^x, b^y)\, b^n \\ \text{"} = \text{"} \quad & S \end{aligned}$$

So, if a probability measure is 'base invariant', it should be invariant at least for the powers of the initial base, and so the probability of S should equal that of $S^{\frac{1}{m}}$. Hence the definition is somehow weak, since it only accounts for the base that are powers of the initial base, but the following theorem 2.3.7 will show that it is sufficient for *any* base.

Now, let P_b denote the logarithmic probability measure on $(\mathbb{R}^+, \mathcal{M}_b)$, i.e.:

$$\forall\, t \in [1,b), P_b \left(\bigcup_{n=-\infty}^{\infty} [1,t]\, b^n \right) = \log_b(t)$$

Let also Δ_1 denote the Dirac delta measure of the set $S_1 = \bigcup_{n=-\infty}^{\infty} \{1\} \times b^n$. More precisely, Δ_1 is defined for all $S \in \mathcal{M}_b$ by $\Delta_1(S) = 1$ if $S \supseteq S_1$ and 0 otherwise (this definition is valid because S_1 has no nonempty $\mathcal{M}_b$-measurable subset).

The following theorem links base invariance with those probability measures, and thus with Benford's law:

Theorem 2.3.7 *A probability measure P on $(\mathbb{R}^+, \mathcal{M}_b)$ is base invariant if there exists $q \in [0,1]$ s.t.*

$$P = (1-q)P_b + q\Delta_1$$

Corollary 2.3.8 *The unique base-invariant and atomless probability measure on* $(\mathbb{R}^+, \mathcal{M}_b)$ *is* P_b.

Before proving that theorem, a few lemmae must first be stated.

Definition 2.3.9 *A measure* μ *on* $(\Omega, \mathcal{F})$ *is invariant under the measure mapping* $T : \Omega \to \Omega$ *if*

$$\mu(E) = \mu(T^{-1}(E)),\ \forall E \in \mathcal{F}$$

.

Now let n be an arbitrary positive integer, and consider the Borel measurable space on $[0,1)$ and the mapping T_n defined on $[0,1)$ by $T_n(x) = nx \mod 1$.

Lemma 2.3.10 *A probability measure* $\overline{P}$ *on* $\big([0,1), \mathbb{B}[0,1]\big)$ *is invariant under* T_n *if and only if*

$$\overline{P}[0,a) = \sum_{k=0}^{n-1} \overline{P}\left[\frac{k}{n}, \frac{k+a}{n}\right)$$

Proof.
It is here sufficient to show that:

$$T_n\left(\bigcup_{k=0}^{n-1}\left[\frac{k}{n}, \frac{k+a}{n}\right)\right) = [0,a)$$

- First, take $x \in \bigcup_{k=0}^{n-1}\left[\frac{k}{n}, \frac{k+a}{n}\right)$.
  $\Rightarrow \exists\, k' \in \{0, \ldots, n-1\}$ s.t. $k' \leq nx < k' + a$
  $\Rightarrow nx \mod 1 \in [0,a)$
- Conversely, take $y \in [0,a)$.
  $\Rightarrow \forall\, k \in \mathbb{N}, \exists\, x' \in [k, k+a)$ s.t. $x' = k + y$
  $\Rightarrow \forall\, k \in \mathbb{N}, \exists\, x \in \left[\frac{k}{n}, \frac{k+a}{n}\right)$ s.t. $nx = k + y$ (or $y = nx \mod 1$)
  Now for x has to be restrained to $[0,1)$, k must be in $\{0, \ldots, n-1\}$.
  $\Rightarrow \forall\, k \in \{0, \ldots, n-1\}, \exists\, x \in \left[\frac{k}{n}, \frac{k+a}{n}\right)$ s.t. $y = nx \mod 1$

■

Now a second lemma can be stated. Let λ denote the Lebesgue measure on $[0,1)$, and δ_0 the Dirac probability measure at 0.

Lemma 2.3.11 *A Borel probability measure* $\overline{P}$ *on* $[0,1)$ *is invariant under* T_n *if and only if*

$$\overline{P} = q\delta_0 + (1-q)\lambda$$

for some $q \in [0,1]$.

Proof.
First, let $\overline{P} = q\delta_0 + (1-q)\lambda$ for some $q \in [0,1]$.
Then for all $a \in [0,1)$:

$$\begin{aligned}
\sum_{k=0}^{n-1} \overline{P}\left[\frac{k}{n}, \frac{k+a}{n}\right) &= q + (1-q)\sum_{k=0}^{n-1} \lambda\left[\frac{k}{n}, \frac{k+a}{n}\right) \\
&= q + (1-q)\sum_{k=0}^{n-1} \frac{a}{n} \\
&= q + (1-q)a \\
&= \overline{P}[0,a)
\end{aligned}$$

Conversely, suppose $\overline{P}$ is an arbitary probability measure on $[0,1)$ invariant under T_n.

Recall that a probability measure on $[0,1)$ is uniquely determined by its Fourier coefficients:

$$\phi_n = \int_0^1 e^{2i\pi nx}\, d\overline{P}(x) \ , \ n \in \mathbb{N}$$

.

The invariance of $\overline{P}$ under T_n implies that its Fourier coefficients ϕ_n are constant, for all $n \in \mathbb{N}$.

To see this, use the change of variable $x' = T_n(x) = nx \mod 1$, and then remark that $e^{2i\pi x'} = e^{2i\pi nx}$:

$$\begin{aligned}
\phi_n &= \int_0^1 e^{2i\pi nx} d\overline{P}(x) \\
&= \int_0^1 e^{2i\pi x'} d\overline{P}(T_n^{-1}(x')) \\
&= \int_0^1 e^{2i\pi x'} d\overline{P}(x') \quad \textit{(since } \overline{P}(E) = \overline{P}(T_n^{-1}(E)) \textit{ for all } E \textit{ in } \mathbb{B}[0,1) \textit{)} \\
&= \phi_1 \quad \text{for all n.}
\end{aligned}$$

Thus let $\phi_n = q$, with $q \in \mathbb{C}$ (*a priori*).
Now, to show that q is in fact a real number in $[0,1)$, consider:

$$\overline{P}(\{0\}) = \int_0^1 \lim_{N\to\infty} \left\{ \frac{1}{N} \sum_{n=1}^{N} e^{2i\pi nx} \right\} d\overline{P}(x)$$

That equality comes from the fact that, given a complex number a with $|a| \leq 1$ (here choose $a = e^{2i\pi x}$), $\frac{1}{N}\sum_{n=1}^{N} a^n \to 0$ as $N \to \infty$, except if $a = 1$ in which case it is equal to 1 for all N.

Now use the bounded convergence theorem (since $\left|\frac{1}{N}\sum_{n=1}^{N} e^{2i\pi nx}\right| \leq 1$) :

$$\begin{aligned}\overline{P}(\{0\}) &= \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\phi_n \\ &= q \quad \textit{(since the } \phi_n \textit{ are constant and equal to } q \textit{)}\end{aligned}$$

Hence q is a probability, and it is in $[0,1)$.
Now, derive the Fourier coefficients ϕ'_n for $q\delta_0 + (1-q)\lambda$:

$$\begin{aligned}\phi'_n &= q \times e^{2i\pi n0} + (1-q)\int_0^1 e^{2i\pi nx}dx \\ &= q \\ &= \phi_n \quad \text{for all n.}\end{aligned}$$

As the Fourier coefficients of a measure determine it uniquely,

$$\overline{P} = q\delta_0 + (1-q)\lambda$$

.

■

Now the proof of the main theorem of base invariance can be easily stated.
Proof of theorem 2.3.7.
First show that P_b and Δ_1 are base invariant.
For Δ_1 this is quite obvious, since the digit 1 is written 1 in any base...
Now for P_b, let x,y be arbitrary numbers in $[0,1)$ and $S = \bigcup_{q=-\infty}^{\infty}[b^x, b^y)\, b^q$:

$$\begin{aligned}P_b(S^{\frac{1}{m}}) &= P_b\left(\bigcup_{q=-\infty}^{\infty}\left(\bigcup_{k=0}^{m-1}[\, b^{\frac{x}{m}}, b^{\frac{y}{m}}\,)\, b^{\frac{k}{m}}\right)\times b^q\right) \\ &\qquad \textit{(using again the proof of property 3 in lemma 2.3.2)} \\ &= \sum_{k=0}^{m-1} P_b\left(\bigcup_{q=-\infty}^{\infty}\left[b^{\frac{x+k}{m}}, b^{\frac{y+k}{m}}\right)\, b^q\right) \\ &= \sum_{k=0}^{m-1}\log_b\left(\frac{b^{\frac{y+k}{m}}}{b^{\frac{x+k}{m}}}\right) \\ &= \sum_{k=0}^{m-1}\log_b\left(b^{\frac{y-x}{m}}\right) \\ &= y-x \\ &= P_b(S)\end{aligned}$$

Hence P_b and Δ_1 are base invariant, and so is their mixture $q\Delta_1 + (1-q)P_b$.

Conversely, suppose P is base invariant on $(\mathbb{R}, \mathcal{M}_b)$.

Let $S = \bigcup_{q=-\infty}^{\infty} [1, b^a)\ b^q$, for some $a \in [0,1)$, and use the expression of $S^{\frac{1}{m}}$ to get:

$$P(S^{\frac{1}{m}}) = P\left(\bigcup_{q=-\infty}^{\infty} \bigcup_{k=0}^{m-1} \left[b^{\frac{k}{m}}, b^{\frac{k+a}{m}} \right)\ b^q \right)$$

Hence, by the measure isomorphism (6), the equality $P(S) = P(S^{\frac{1}{m}})$ implies:

$$\hat{P}[1, b^a) = \hat{P}\left(\bigcup_{k=0}^{m-1} \left[b^{\frac{k}{m}}, b^{\frac{k+a}{m}} \right) \right)$$

And consequently:

$$\overline{P}[0, a) = \sum_{k=0}^{n-1} \overline{P}\left[\frac{k}{n}, \frac{k+a}{n} \right)$$

This last equality means by lemmae 2.3.10 and 2.3.11 that there exists $q \in [0,1)$ such that $\overline{P} = q\delta_0 + (1-q)\lambda$.

Now it can be easily seen that $\overline{\Delta_1} = \delta_0$ and $\overline{P_b} = \lambda$.

Hence there exists $q \in [0,1]$ such that $P = q\Delta_1 + (1-q)P_b$.

■

Remark 2.3.12 *The proof of theorem 2.3.7 explains why the hypothesis "$P(S) = P(S^{\frac{1}{m}})$" is sufficient for invariance in any base (a priori it only implies invariance for powers of the initial base). This condition is in fact sufficient to imply that the probability $\overline{P}$ is $q\delta_0 + (1-q)\lambda$, for some q in $[0,1]$, which in turn implies that $P = q\Delta_1 + (1-q)P_b$ in $(\mathbb{R}, M_b)$, whatever the base b is. Hence P is invariant for any base.*

Remark 2.3.13 *An immediate corollary to theorems 2.3.7 and 2.3.4 is that scale invariance implies base invariance, but not conversely. For example, Δ_1 is base invariant but not scale invariant.*

2.3.4 Random distributions

This section is a summary of [Hilla], where an interesting and modern point of view about Benford's Law is developed.

As was explained in section 1.4, the basic idea of this explanation of Benford's Law is to randomize the distributions themselves. Actually, Benford's experiment is like a collection of random samples from random distributions.

The concept of random distributions has already been studied by several authors, see [Hilla] and the papers referenced there for a complete view. Here only the useful definitions and results in the context will be presented.

Definition 2.3.14 *A (real Borel)* **random probability measure** *(r.p.m)* $\mathbb{M}$ *is a random variable [on an underlying probability space* $(\Omega, \mathcal{F}, \mathbf{P})$*, the space of distributions] that takes value which are Borel probability measures on* $\mathbb{R}$ *and which is regular in the sense that for each Borel set* $B \subset \mathbb{R}$*,* $\mathbb{M}(B)$ *is a random variable.*

Definition 2.3.15 *The* **expected distribution measure** *of a r.p.m* $\mathbb{M}$ *is the probability measure* $\mathbf{E}\mathbb{M}$ *on the Borel measurable space* $(\mathbb{R}, \mathbb{B}(R))$ *defined by:*

$$(\mathbf{E}\mathbb{M})(B) = \mathbf{E}(\mathbb{M}(B)) \text{ for all Borel } B \subset \mathbb{R}$$

where $\mathbf{E}$ *denotes the expectation with respect to* $\mathbf{P}$*.*

The definitions 2.3.14 and 2.3.15 may be a bit abstract, but the underlying process is really simple to understand: the expected distribution of a r.p.m. is a mixture of distributions with respect to a certain probability measure $\mathbf{P}$. For example, if the space of distributions is discrete, consisting of n distributions with densities $f_1, f_2, \ldots f_n$, and the probability $\mathbf{P}$ is uniform on that space, then the expected distribution of the r.p.m. has simply the density $\frac{1}{n}\sum_{i=1}^{n} f_i$.

Now, in order to model Benford's experiment, the concept of $\mathbb{M}$-*random k-sample* is needed.

Definition 2.3.16 *For an r.p.m.* $\mathbb{M}$ *and positive integer k, a* **sequence of** $\mathbb{M}$**-random k-samples** *is a sequence of random variables* $X_1, X_2, \ldots$ *on* $(\Omega, \mathcal{F}, \mathbf{P})$ *so that for some i.i.d. sequences* $\mathbb{M}_1, \mathbb{M}_2, \ldots$ *of r.p.m.'s with the same distribution as* $\mathbb{M}$ *and for each* $j = 1, 2, \ldots,$

given $\mathbb{M}_j = P$*, the random variables* $X_{(j-1)k+1}, \ldots, X_{jk}$ *are i.i.d. with p.d.f.* P

and

$X_{(j-1)k+1}, \ldots, X_{jk}$ *are independent of* $\mathbb{M}_i$ *and* $X_{(i-1)k+1}, \ldots, X_{ik}$ *for all* $i \neq j$*.*

Again definition 2.3.16 may appear as obscure, but the definition is in fact really simple. A sequence of $\mathbb{M}$-random k-samples is concretely a collection of (independent) samples of size k, each sample coming from a distribution "drawn at random and independently" using the r.p.m. $\mathbb{M}$. Hence Benford's experiment is actually a sequence of $\mathbb{M}$-random k-samples, although the r.p.m. used cannot be clearly defined.

However, the following lemma allows to go on in the analysis, even if the r.p.m. is not completely known:

Lemma 2.3.17 *Let* $\mathbb{M}$ *be a r.p.m., and let* $X_1, X_2, \ldots$ *be a sequence of* $\mathbb{M}$*-random k-samples for some k. Then for any Borel B,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \in B} = \mathbf{E}\mathbb{M}(B)$$

Lemma 2.3.17 means that, even though the X_i are not strictly i.i.d, the observed frequency of an event in a sequence of $\mathbb{M}$-random k-samples still converges to a probability, and this probability is the measure of that event according to the expected distribution of $\mathbb{M}$. So in one word to know the frequency of an event only requires to know the expected distribution.

A detailed explanation of why the X_i are not i.i.d, and also the proof of lemma 2.3.17, can be found in [Hilla].

Now the background is set for the next result, which is a 'short' version of Hill's theorem.

Theorem 2.3.18 (Log-limit law for significant digits) *Let $\mathbb{M}$ be an r.p.m. on $(\mathbb{R}^+, \mathcal{M})$. The following are equivalent:*

1. $\mathbf{E}\mathbb{M}$ *is scale-invariant;*
2. $\mathbf{E}\mathbb{M}$ *is base-invariant and atomless on $(\mathbb{R}^+, \mathcal{M})$;*
3. $\mathbf{E}\mathbb{M}\left(\bigcup_{n=-\infty}^{\infty}[1,t)\,10^n\right) = \log(t)$ *for all $t \in [1,10)$;*
4. *for every sequence of $\mathbb{M}$-random k-sample $X_1, X_2 \ldots$,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} 1_{M(X_i)\in[1,t)} = log(t) \text{ for all } t \in [1,10)$$

.

Proof. The proof is immediate:
(1) $\Leftrightarrow$ (3) by theorem 2.3.4.
(2) $\Leftrightarrow$ (3) by corollary 2.3.8 of theorem 2.3.7.
(3) $\Leftrightarrow$ (4) by lemma 2.3.17.
■

This theorem explains why Benford's Law may appear in such an experiment as Benford did. Observing Benford's Law does not require that *individual* realization of $\mathbb{M}$ are scale- or base-invariant, only **on average** has it to be. In Benford's experiment some of the distributions were apparently far from satisfying Benford's Law (certain mathematical sequences for instance), but combining them with others seems to have created a scale- and/or base-invariant average distribution.

In a sequence of $\mathbb{M}$-random k-samples, only the expected distribution plays a role in frequency calculations, so inferring on that distribution allows one to draw conclusions on frequencies. Thus the average distribution $\mathbf{E}\mathbb{M}$ is a sort of practical tool, and all the questions about the complete "behind-the-scene" space of distributions are avoided. In that context the space of distributions is a bit like the formerly used "universal table of constants"... Hence in a sense the problem remains open, even if it can't be denied that Hill's approach explains why

> there are many (natural) sampling procedures which lead to the log distribution.

Remark 2.3.19 *Theorem 2.3.18 can be extended to cases where the size of the samples k is not fixed (random for example)*

2.4 An invariant-sum characterization

This section aims at exploring another characterization of Benford's law, which was first observed by Nigrini and later proved by Allaart, see [Allaart].

2.4.1 Sum-invariance

In his Ph.D. thesis (1992), Nigrini observed that tables of unmanipulated accounting data closely satisfy Benford's Law, and that in such lists of data,

> The sum of all entries with leading digit d is constant for various d.

Here the word "entries" means "mantissae". Of course if the numbers themselves were summed, the sum would not be constant, as for example a peculiar large number would dominate in a sum.

In fact it makes sense when one looks at the distribution. For instance in a sample of integers that follows a first-digit logarithmic law, there are lots of 1, a bit less of 2, ... and so on until 9. So the sum of 1s should be "roughly" (i.e. in expectation) equal to that of 2s, to that of 3s, etc...

Now what was said for the first digit was also observed for the two leading digits, the three leading digits, etc... For example the sum of all the entries that begin with 1.234 is equal (in expectation) to the sum of all the entries that begin with 7.453... (note that the considered number of significant digits has to be the same, here 4 for example). So intuitively a distribution can be qualified as sum-invariant if for any natural number k (fixed), the expected sum of the mantissae of all entries starting with a certain k-tuple of significant digits is the same as that for any other k-tuple.

2.4.2 Equivalence between sum-invariance and logarithmic law

Allaart has shown that the sum-invariance property is a characterization of the logarithmic law, as no other law has that property. First a few notations have to be given:

Given $k \in \mathbb{N}$, $d_1 \in \{1, 2, \dots, 9\}$ and $d_2, d_3, \dots d_k \in \{0, 1, \dots, 9\}$:

- let $A(d_1, d_2, \dots d_k)$ be the set of real numbers whose first significant digits are $d_1, d_2, \dots d_k$.
- let $\overline{A}(d_1, d_2, \dots d_k)$ be the set of real numbers in $[1, 10)$ whose first significant digits are $d_1, d_2, \dots d_k$.

Given a probability measure P on $(\mathbb{R}^+, \mathbb{B}(\mathbb{R}))$, define P_M on $([1,10), \mathbb{B}([1,10))$ as the corresponding mantissa probability measure, i.e.:

$$P_M[1,t) = P\left(\bigcup_{n=-\infty}^{\infty} [1,t)\,10^n\right), \quad \forall\, t \in [1,10)$$

(P_M is in fact the same as what was called $\hat{P}$ in section 2.3)

It has now to be remarked that in the so called sum-invariance property the summation is not necessary, only the expectation is. The summation as Nigrini did is in fact a way to simulate the average. Take X an arbitrary random variable that comes from a sum-invariant distribution, and let $k \in \mathbb{N}$. What is really meant by sum-invariance is that the (conditional) expectation of $M(X)$ given that $M(X)$ begins with $d_1, d_2, \ldots, d_k$ is the same whatever the k-tuple $(d_1, d_2, \ldots, d_k)$ is. Hence a formal definition of sum-invariance can be given as follow:

Definition 2.4.1 *A probability measure P on $(\mathbb{R}^+, \mathbb{B}(\mathbb{R}^+))$ is sum-invariant if, for any random variable X with distribution P, and for any fixed $k \in \mathbb{N}$, the expectations $E[M(X)1_{A(d_1,\ldots,d_k)}(X)]$ are constant whatever the k-tuple $(d_1, d_2, \ldots, d_k)$ is (with $d_1 \in \{1,2,\ldots,9\}$ and $d_2, d_3, \ldots d_k \in \{0,1,\ldots,9\}$).*

Now the main theorem can be stated:

Theorem 2.4.2 *A probability measure P on $(\mathbb{R}^+, \mathbb{B}(\mathbb{R}^+))$ is sum-invariant if and only if $P_M[1,t) = log(t)$.*

Proof.

First observe that, for all digits $d_1, d_2, \ldots d_k$:

$$E[M(X)1_{A(d_1,\ldots,d_k)}(X)] = E[M(X)1_{\overline{A}(d_1,\ldots,d_k)}(X)] = \int_{\overline{A}(d_1,\ldots,d_k)} x\, dP_M(x) \tag{7}$$

Then remark that all the sets $\overline{A}(d_1, \ldots, d_k)$ are in fact intervals, have the same length, say $\lambda(A)$ *(here λ denotes the Lebesgue measure on $[1,10)$)*, and form a partition of $[1,10)$.

Suppose sum-invariance, i.e. the expectations in (7) are constant (denote their value $\int_A x\, dP_M(x)$), and sum them over all sets $\overline{A}(d_1, \ldots, d_k)$:

$$\sum_{\substack{1\le d_1\le 9\\ 0\le d_2\le 9\\ \cdots\\ 0\le d_k\le 9}} \int_{\overline{A}(d_1,\ldots,d_k)} x\, dP_M(x) = \int_1^{10} x\, dP_M(x)$$

$$\Rightarrow \frac{9}{\lambda(A)} \int_A x\, dP_M(x) = \int_1^{10} x\, dP_M(x)$$

(since the expectations are constant and that there are $\frac{9}{\lambda(A)}$ partitioning intervals of length $\lambda(A)$ in $[1,10)$)

$$\Rightarrow \int_A x\, dP_M(x) = \frac{\lambda(A)}{9} \int_1^{10} x\, dP_M(x) \tag{8}$$

Conversely, if the equality (8) is valid for all sets A of the form $\overline{A}(d_1, \ldots, d_k)$, then P is obviously sum-invariant since then the expectations in (7) are constant.

Hence the equality (8) is a necessary and sufficient condition for sum-invariance.

Now suppose P_M is the logarithmic law $\left(dP_M(x) = \frac{1}{x \ln 10} dx\right)$, and substitute in both sides of (8):

$$\int_A x\, dP_M(x) = \frac{1}{\ln 10} \lambda(A)$$

and:

$$\frac{\lambda(A)}{9} \int_1^{10} x\, dP_M(x) = \frac{\lambda(A)}{9} \times \frac{9}{\ln 10}$$

So the equality (8) is verified.

Conversely, suppose that (8) holds for all A of the form $\overline{A}(d_1, \ldots, d_k)$. Every interval of $[1,10)$ can be represented as a countable union of such sets, so by summing the integrals, (8) holds for every interval.

Both sides of (8) can define a probability measure of the set A, and those measures are thus equal on the set of intervals of $[1,10)$. By Caratheodory's extension theorem, this is enough to ensure their complete equality on the set of Borels of $[1,10)$.

Hence $x\, dP_M(x)$ is proportional to dx, and so $dP_M(x)$ is proportional to $\frac{1}{x} dx$.

By normalizing over $[1,10)$ it follows that $dP_M(x) = \frac{1}{x \ln 10} dx$, i.e. P_M is the logarithmic law.

■

2.5 Other invariances

This section briefly presents some other interesting properties of Benford's Law, which may explain why it occurs so frequently, especially in computing arithmetics.

2.5.1 Inverse

If X satisfies Benford's Law, quite a few basic functions of X were found to satisfy Benford's Law as well. Maybe the most simple is the inverse (this property is taken from [Adhikari & Sarkar]).

Proposition 2.5.1 *If X is a random variable such that $M(X)$ follows the logarithmic law, then $M(X^{-1})$ also follows the logarithmic law.*

Proof. For all $t \in [1, 10)$:

$$\begin{aligned}
\Pr(M(X^{-1}) \in [1,t)) &= \Pr\left(X^{-1} \in \bigcup_{n=-\infty}^{\infty} [1,t)\, 10^n\right) \\
&= \Pr\left(X \in \bigcup_{n=-\infty}^{\infty} \left[\frac{1}{t}, 1\right)\, 10^n\right) \\
&= \Pr\left(M(X) \in \left[\frac{10}{t}, 10\right)\right) \\
&= 1 - \log\left(\frac{10}{t}\right) \\
&= \log(t)
\end{aligned}$$

■

2.5.2 Multiplication and division

This part is inspired from [Hamming]. Because the exponents do not play a role in the mantissa of the product of two real numbers, it is rather easy to derive an expression of its distribution.

Proposition 2.5.2 *Let X, Y be two random variables, and let f, g and h be the respective densities of $M_b(X)$, $M_b(Y)$, and $M_b(XY)$. Then, for all $z \in [1, b)$:*

$$h(z) = \int_1^z \frac{1}{x}\, g\left(\frac{z}{x}\right) f(x)\, dx + \int_z^b \frac{b}{x}\, g\left(\frac{bz}{x}\right) f(x)\, dx \tag{9}$$

Proof. Let F, G, and H be the respective probability distribution functions of $M_b(X)$, $M_b(Y)$, and $M_b(XY)$.

Now let x and y be two real numbers. Without loss of generality, suppose x, y in $[1, b)$. The mantissa of xy is then given by:

$$M_b(xy) = \begin{cases} xy \text{ if } 1 \leq xy < b \\ \\ \frac{xy}{b} \text{ if } b \leq xy < b^2 \end{cases}$$

Hence for all $z \in [1, b)$,

$$\begin{aligned}
M_b(xy) \leq z \iff & \begin{cases} xy \leq z \\ \text{or} \\ \frac{xy}{b} \leq z \text{ and } b \leq xy \end{cases} \\
\iff & \begin{cases} y \leq \frac{z}{x} \\ \text{or} \\ \frac{b}{x} \leq y \leq \frac{zb}{x} \end{cases}
\end{aligned}$$

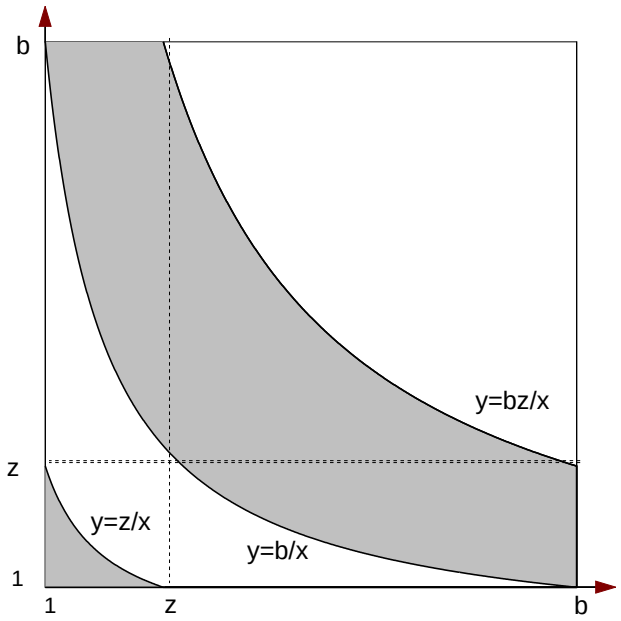


Figure 2.3: $M_b(xy) \leq z$

Divide the shaded region into three and integrate:

$$\begin{aligned} H(z) &= \int_1^z \int_1^{\frac{z}{x}} f(x)g(y)\,dydx + \int_1^z \int_{\frac{b}{x}}^b f(x)g(y)\,dydx + \int_z^b \int_{\frac{b}{x}}^{\frac{zb}{x}} f(x)g(y)\,dydx \\ &= \int_1^z \left[G\left(\frac{z}{x}\right) - G(1) + G(b) - G\left(\frac{b}{x}\right)\right] f(x)\,dx \\ &\quad + \int_z^b \left[G\left(\frac{bz}{x}\right) - G\left(\frac{b}{x}\right)\right] f(x)\,dx \end{aligned}$$

Differentiate with respect to z:

$$\begin{aligned} h(z) &= f(z)\left[G(1) - G(1) + G(b) - G\left(\frac{b}{z}\right) - G(b) + G\left(\frac{b}{z}\right)\right] \\ &\quad + \int_1^z \frac{1}{x}\, g\left(\frac{z}{x}\right) f(x)\,dx + \int_z^b \frac{b}{x}\, g\left(\frac{bz}{x}\right) f(x)\,dx \\ &= \int_1^z \frac{1}{x}\, g\left(\frac{z}{x}\right) f(x)\,dx + \int_z^b \frac{b}{x}\, g\left(\frac{bz}{x}\right) f(x)\,dx \end{aligned}$$

■

Invariance by multiplication is a quite amazing property. In fact it is true for multiplication by *any* (continuous) random variable:

Proposition 2.5.3 *Let Y be a random variable that satisfies Benford's Law. Let X be an arbitrary random variable with a continuous density. Then XY satisfies Benford's Law.*

Proof. With the same notations as in proposition 2.5.2, let $g(y) = \frac{1}{y \ln b}$ in (9) (i.e. Y satisfies Benford's Law):

$$\begin{aligned} h(z) &= \int_1^z \frac{1}{x} \frac{x}{z \ln b} f(x)\, dx + \int_z^b \frac{b}{x} \frac{x}{bz \ln b} f(x)\, dx \\ &= \frac{1}{z \ln b} \int_1^b f(x)\, dx \\ &= \frac{1}{z \ln b} \quad \text{(i.e. XY satisfies Benford's Law)} \end{aligned} \tag{10}$$

■

An immediate corollary of propositions 2.5.1 and 2.5.3 is that invariance is also true for division:

Proposition 2.5.4 *Let Y be a (nonzero) random variable that satisfies Benford's Law. Let X be an arbitrary (nonzero) random variable with a continuous density. Then $\frac{X}{Y}$ and $\frac{Y}{X}$ satisfy Benford's Law.*

These results (propositions 2.5.3 and 2.5.4) guarantee that, in the case of a long series of multiplication and/or divisions, if one of the operand happens by any mean to satisfy Benford's Law, then the whole following sequence of numbers will satisfy Benford's Law. The only drawback of this is that one operand has to satisfy "exactly" Benford's Law.

2.5.3 Convergence

Besides the results presented in section 2.5.2, Hamming found there is an even more amazing result about multiplication, which explains why in long computations Benford's Law might appear "out of nowhere". In fact a product of random variables happens to be "closer" (in the following sense) to Benford's Law than its operands.

Definition 2.5.5 *Define the (relative) distance of density f to Benford's Law by:*

$$D(f) = \max_{1 \le z < b} \left| \frac{f(z) - r(z)}{r(z)} \right|$$

where $r(z) = \frac{1}{z \ln b}$.

Now for the announced (very simple) result:

Proposition 2.5.6 *[With the same notations as in proposition 2.5.2]*

$$D(h) \leq \min(D(f), D(g))$$

Proof.
Equation (10) says that, for any density f:

$$r(z) = \int_1^z \frac{1}{x}\, r\left(\frac{z}{x}\right)\, f(x)\, dx + \int_z^b \frac{b}{x}\, r\left(\frac{bz}{x}\right)\, f(x)\, dx$$

[Note that this can be found directly by a change of variable]
Now substract this from expression (9), and divide by $r(z)$:

$$\begin{aligned}
\frac{h(z) - r(z)}{r(z)} &= \int_1^z \frac{f(x)}{x}\, \frac{g\left(\frac{z}{x}\right) - r\left(\frac{z}{x}\right)}{r(z)}\, dx \\
&\quad + \int_z^b \frac{bf(x)}{x}\, \frac{g\left(\frac{bz}{x}\right) - r\left(\frac{bz}{x}\right)}{r(z)}\, dx \\
&= \int_1^z f(x)\, \frac{g\left(\frac{z}{x}\right) - r\left(\frac{z}{x}\right)}{r\left(\frac{z}{x}\right)}\, dx \\
&\quad + \int_z^b f(x)\, \frac{g\left(\frac{bz}{x}\right) - r\left(\frac{bz}{x}\right)}{r\left(\frac{bz}{x}\right)}\, dx
\end{aligned}$$

Since $f(x) \geq 0$ on $[1, b)$,

$$\begin{aligned}
\left|\frac{h(z) - r(z)}{r(z)}\right| &\leq \int_1^z f(x) D(g)\, dx + \int_z^b f(x) D(g)\, dx \\
&\leq D(g)
\end{aligned}$$

Therefore $D(h) \leq D(g)$, and since f and g are clearly interchangeable, $D(h) \leq \min(D(f), D(g))$.

Of course this doesn't mean there is an actual convergence, since for example if f is the Dirac function δ_1 then $D(h) = D(g)$. But in many practical cases the inequality will be strict so the convergence will happen (see for instance [Schatte] in which various cases are reviewed, and see also the case of a product of uniforms below).

■

Remark 2.5.7 *The result 2.5.6 is also true for division, but proving it requires an expression of the density of the mantissa of a ratio -not difficult but a bit long- (c.f. [Hamming] again).*

Adhikari and Sarkar have found that, if $U_i, i \in \mathbb{N}$ are uniform independent random variables, asymptotically the sequences $U_1 U_2 \ldots U_n$ and U_1^n satisfy Benford's Law. This is of course related to the result 2.5.6, which is however more general.

Actually, it is possible to get an exact expression of the distance of a product $U_1 U_2 \dots U_n$ to Benford's Law, according to the number of operands n. In [Hamming] the following results are presented:

Number of operands	Distance to Benford's Law
1	1.558
2	0.3454
3	0.0980
4	0.0289

This shows that convergence, in the case of a product of uniforms, takes place rather quickly...

2.5.4 Addition and subtraction

The principle of invariance described in 2.5.2 also exists for addition, but it is not as straigthforward. In fact there is no easy formula for the mantissa of a sum, as the exponents of the operands also play a big part in the result.

Hamming has proved that, under certain suppositions, addition leaves Benford's Law invariant (see [Hamming]). Moreover some authors have shown that the distribution of a sum of random variables converges to Benford's Law, in a certain sense (H_∞ and Riesz means, see [Schatte] and the papers referenced there). Of course there are lots of summability methods, and everything can be meant by "a certain sense", so this part alone would require a very long study...

Even if invariance by addition leaves the author a bit unconvinced, it was found that sometimes empirically it works quite well (see section 3.1.2).

2.6 Sequences and Benford's law

Since Benford used mathematical sequences in his experiment, and also since he tried to prove his law on the natural sequence of integers, many authors were interested in investigating Benford's Law for sequences.

2.6.1 Definition and useful theorems

In what follows the base considered is 10 for simplicity, but as always everything can be easily extended to any base. For sequences Benford's Law has a peculiar expression, i.e. the definition of *Benford sequences*:

Definition 2.6.1 *A real sequence $\{a_n\}_{n\in\mathbb{N}}$ is called a Benford sequence if:*

$$\lim_{N\to\infty} \# \left\{ \frac{1}{N} \sum_{n=1}^{N} 1_{M(a_n)\leq t} \right\} = \log t$$

Remark 2.6.2 *The definition 2.6.1 is a classical one, it only means that the limiting frequency tends to a logarithmic law. That is why such sequences are sometimes called "strong" Benford sequences. Other definitions have been given (see for example [Raimi]), and they lead to "weak" Benford sequences. But again many such definitions can exist, so they will not be considered here.*

Benford sequences are closely related to the famous *uniformly distributed modulo 1* sequences, first studied by Hermann Weil in 1916 (see [Weyl]). All the known results about these sequences are summed up in [Kuipers & Niederreiter].

The following theorem (see for example [Diaconis]) makes the relation explicit:

Theorem 2.6.3 *The real sequence $\{a_n\}_{n\in\mathbb{N}}$ is a Benford sequence if and only if the sequence $\{\log(a_n)\}_{n\in\mathbb{N}}$ is uniformly distributed modulo 1.*

Proof. The proof given here is a short (and maybe not so rigorous) one, see again [Diaconis] for a detailed one.

If $\{\log(a_n)\}$ is uniformly distributed modulo 1, this sequence can be considered as a sample from a uniform distribution. According to the variate generation procedure (section 2.1.4, and remark 2.1.10), the sequence $\{a_n\}$ can thus be considered as a sample from a distribution that satisfies Benford's Law. Because of the frequency-oriented definition 2.6.1, $\{a_n\}$ is hence a Benford sequence (and conversely).

■

Now the following useful theorems about uniformly distributed sequences (modulo 1) have to be presented (see chapters 1 and 2 of [Kuipers & Niederreiter] for the proofs):

Theorem 2.6.4 *If $\{x_n\}$ is uniformly distributed modulo 1, and $\{y_n\}$ is such that $\lim(x_n - y_n) = \alpha \in \mathbb{R}$ is constant, then $\{y_n\}$ is uniformly distributed modulo 1.*

Theorem 2.6.5 (Weyl's criterion) *$\{x_n\}$ is uniformly distributed modulo 1 if and only if, for all $h \in \mathbb{N}^*$,*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} e^{2i\pi h x_n} = 0$$

.

Theorem 2.6.6 (Van der Corput's estimate for trigonometric sums) *Let $a, b \in \mathbb{Z}$, $a < b$, and f a function twice differentiable on $[a, b]$ with $f''(x) \geq \rho > 0$ or $f''(x) \leq -\rho < 0$ for $x \in [a, b]$. Then*

$$\left| \sum_{n=a}^{b} e^{2i\pi f(n)} \right| \leq (|f'(a) - f'(b)| + 2) \left(\frac{4}{\rho} + 3 \right) \tag{11}$$

Theorem 2.6.7 (Fejer's theorem) *If $\{f(n)\}$ is uniformly distributed modulo 1, then*

$$\limsup\ n|f(n+1) - f(n)| = \infty$$

.

2.6.2 Geometric sequences

Proposition 2.6.8 *A geometric sequence* $\{a^n\}$ *is a Benford sequence if and only if* $\log(a)$ *is irrational.*

Proof. By theorem 2.6.3, $\{a^n\}$ is a Benford sequence if and only if $\{n\log(a)\}$ is uniformly distributed mod 1. It has long been known that the only sequences of the form $\{n\alpha\}$ that are uniformly distributed modulo 1 are those for which α is irrational (see [Weyl] again).

■

Example 2.6.9 *Of course this case regroups* **lots** *of classical sequences, among which the basic* $\{2^n\}$*... This only can explain many observations of Benford's Law in nature.*

2.6.3 $\{n!\}$ and $\{n^n\}$

The two sequences $\{n!\}$ and $\{n^n\}$ are Benford sequences. They are included here to show some examples of the joint use of Weyl's fundamental criterion (theorem 2.6.5) and Van der Corput's estimate (theorem 2.6.6).

- For $\{n^n\}$ this is quite straightforward:

  Let $h \in \mathbb{Z}^*$, $N \in \mathbb{N}^*$, and put $a = 1$, $b = N$, and $f(n) = hn\log(n)$ in (11).

$$\begin{aligned} f'(n) &= h\log(n) + h \\ f''(n) &= \frac{h}{n} \quad \left(\Rightarrow \rho = \frac{|h|}{n}\right) \end{aligned}$$

$$\begin{aligned} \Rightarrow \frac{1}{N}\left|\sum_{n=1}^{N} e^{2i\pi h n\log(n)}\right| &\leq \frac{1}{N}(|h|\log(N)+2)\left(4\sqrt{\frac{N}{|h|}}+3\right) \\ &= O\left(\frac{\log(N)}{N^{\frac{1}{2}}}\right) \\ &\to 0 \quad \text{as } N \to \infty \end{aligned}$$

  Hence by Weyl's criterion $\{n\log(n)\}$ is uniformly distributed mod 1, and by theorem 2.6.3 $\{n^n\}$ is a Benford sequence.

- For $\{n!\}$ this is a bit more difficult:

  First recall Stirling's formula:

$$n! \sim \frac{1}{\sqrt{2\pi}}\, n^{\left(n+\frac{1}{2}\right)} e^{-n}$$

  .

Thus the sequence

$$\log(n!) - \left(\left(n+\frac{1}{2}\right)\log(n) - \frac{1}{\ln(10)}n\right)$$

tends to a constant as $n \to \infty$, and by theorems 2.6.4 it is hence sufficient to show that $\left\{\left(n+\frac{1}{2}\right)\log(n)+kn\right\}$ is uniformly distributed modulo 1 $\left(\text{where } k = -\frac{1}{\ln(10)} \text{ constant}\right)$.

So let $h \in \mathbb{Z}^*$, $N \in \mathbb{N}^*$, and put $a = 1, b = N$, and $f(n) = h\left(n+\frac{1}{2}\right)\log(n) + hkn$ in (11).

$$\begin{aligned} f'(n) &= h\log(n) + h + \frac{h}{2n} + hk \\ f''(n) &= \frac{h}{n} - \frac{h}{2n^2} \quad \left(\Rightarrow \rho = |h|\left(\frac{1}{N} - \frac{1}{2N^2}\right)\right) \end{aligned}$$

$$\begin{aligned} \Rightarrow \frac{1}{N}\left|\sum_{n=1}^{N} e^{2i\pi f(n)}\right| &\leq \frac{1}{N}\left(|h|\log(N) + \left|\frac{h}{2}\right|\left(\frac{1}{N}-1\right)+2\right)\left(\frac{4\sqrt{2}N}{\sqrt{|h|(2N-1)}}+3\right) \\ &= O\left(\frac{\log(N)}{N^{\frac{1}{2}}}\right) \\ &\to 0 \quad \text{as } N \to \infty \end{aligned}$$

Finally the use of Weyl's criterion and theorem 2.6.3 completes the proof.

2.6.4 Other sequences

Here some non-Benford sequences will be presented. In proving that a sequence is not Benford, theorem 2.6.7 is very useful, or rather its following immediate corollary (use theorem 2.6.3 and $f(n) = \ln(a_n)$):

Corollary 2.6.10 *If $\{a_n\}$ is a Benford sequence, then $n\ln\left(\frac{a_{n+1}}{a_n}\right) \to \infty$.*

Using that corollary, $\{n^b\}$, $\{bn\}$ and $\{\log_b n\}$ can be easily proved not to be Benford, whatever b is.

- For $\{n^b\}$: $n\ln\left(\frac{(n+1)^b}{n^b}\right) = nb\ln\left(1+\frac{1}{n}\right) \to b$
- For $\{bn\}$: $n\ln\left(\frac{b(n+1)}{bn}\right) = n\ln\left(1+\frac{1}{n}\right) \to 1$
- For $\{log_b n\}$: $n\ln\left(\frac{\log_b(n+1)}{\log_b n}\right) = n\ln\left(1+\frac{\ln\left(\frac{n+1}{n}\right)}{\ln n}\right) \sim \frac{1}{\ln n} \to 0$

And, finally, to be complete, several other sequences were found to be Benford, like $\left\{\binom{n}{k}\right\}$ (see [Diaconis]), the Lucas and Fibonacci sequences, and other recursively defined sequences (see [Schatte] and the papers referenced there)...

Part II

Experiments on Benford's Law

Chapter 3

Real datasets and Benford's Law

3.1 U.S. counties and towns

This section presents results about some demographic data coming from the U.S. Census Bureau (www.census.gov).

3.1.1 Populations of the U.S. counties

In 1995 Nigrini and Wood found that the 1990 census populations of the 3141 counties in the United States followed Benford's law (see [Hilla]). They also found that it was true for the predicted values for 1991 and 1992.

10 years after, in 2000, another census was organized in the U.S., so the idea of checking Benford's Law on that available data was quite tantalizing. If it was true, then the "Benford-in-Benford-out" test (see section 1.6) would effectively be useful in building models for population growth.

Figures 3.1 and 3.2 present histograms of the first digit for both 2000 and 1990.

Apparently, both first digits are very 'close' to Benford's Law. Quantifying the distance is however a bit tricky, as in samples of such size the Chi-square test always rejects the null hypothesis for reasonable confidence levels.

Recall that the Chi-square statistic s is defined by:

$$s = N \times \sum_{i=1}^{n} \frac{(f_i - e_i)^2}{e_i}$$

where N is the sample size, n is the number of classes, and f_i, e_i are respectively the observed and expected proportions for the class i.

In the Chi-square test, the null hypothesis is "the sample comes from the distribution with p.d.f. $\{f_i\}_{i=1..n}$", so the null hypothesis is rejected if s is large.

Recall also that under the null hypothesis $S \sim \chi^2_{n-1}$. Thus the p-value for the test is $1 - F_{n-1}(s)$, where F_{n-1} is the c.d.f. of χ^2_{n-1}.

However if N is large, s will tend to be large, and even if there are many entries, the sample may not be able to approach enough the expected law to decrease s. This can be seen as a limit of the Neyman-Pearson theory, see [Ley] for instance and the papers referenced there.

That is why, here and throughout, both the normal Chi-square and the distance $\sum_{i=1}^{n}(f_i - e_i)^2$ (with the same notations as in the Chi-square) will be shown. The analysis then requires more 'common sense' than strict mathematical calculations. The author finds that the decimal order of the distance provides a good idea of how Benford's Law is fit by a dataset. In particular, for the first digit in base 10, a distance of the order of 10^{-4} can be considered as a medium value for conformity: it means that on average $|f_i - e_i|$ is in the order of between 10^{-2} and 10^{-3} (i.e. the absolute difference between the observed frequency and the predicted probability is on average between 1% and 0.1%) . The Chi-square and the p-value must be considered with a lot of tolerance.

For the first digit of the 3141 U.S. counties the results are the following:

Year	Chi-square	Distance	p-value
1990	15.2	$6.2\ 10^{-4}$	5%
2000	10.0	$3.0\ 10^{-4}$	26%

Now what could be the explanation of such a phenomenon?

There might be a very simple answer. In fact it makes sense to suppose that the population of a given county is a geometric sequence with respect to time, since its growth rate is roughly constant. Hence it follows Benford's Law (see section 2.6.2) over time. And now if one takes lots of counties at **one instant** (like in a survey), there should be counties at every "stage" of time. In the language of time series, the process of surveying seems to keep Benford's Law invariant from time distribution to ensemble distribution.

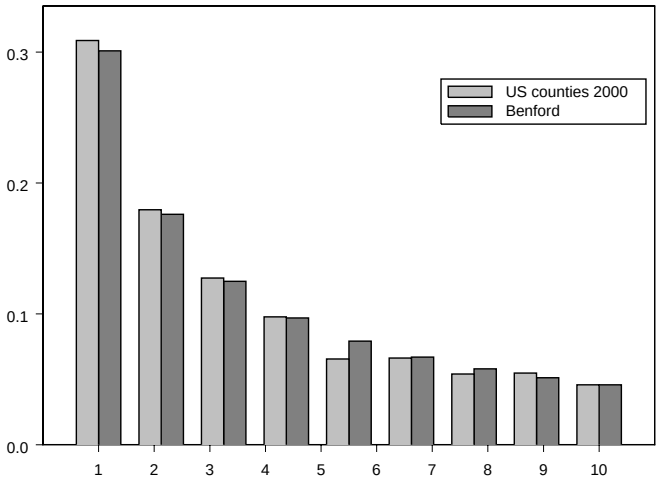


Figure 3.1: Population of the 3141 U.S. counties in 2000

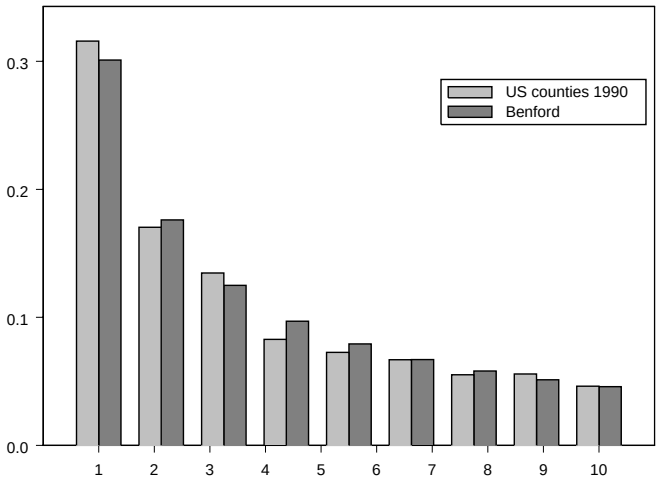


Figure 3.2: Population of the 3141 U.S. counties in 1990

3.1.2 Absolute and relative changes

The absolute change is the net change of populations in the U.S. counties between 1990 and 2000. The change is taken as an absolute value since some counties have lost inhabitants.

The relative change is simply the change in percentage, once again in absolute value.

Figure 3.3 and 3.4 show that absolute change is quite close from Benford's Law, whereas relative change is not really. What somehow confirms this observation are the Chi-square results:

Type	Chi-square	Distance	p-value
Absolute	9.6	$2.8\ 10^{-4}$	30%
Relative	66.9	$3.0\ 10^{-3}$	0%

An explanation of Benford's Law is here a bit obscure. Absolute change can be seen as a realization of $X - Y$, where both X and Y satisfy Benford's Law... Therefore it allows to think that invariance by substraction (and by addition) is 'sometimes' true.

Relative change raises another problem. It can be seen as a realization of $\frac{X-Y}{X}$, so by multiplication invariance (see section 2.5.2), it would normally be a good fit to Benford's Law. But there the limit of observation by 'common sense' is reached, i.e. the question "is the fit on figure 3.4 good or not?" cannot be precisely answered.

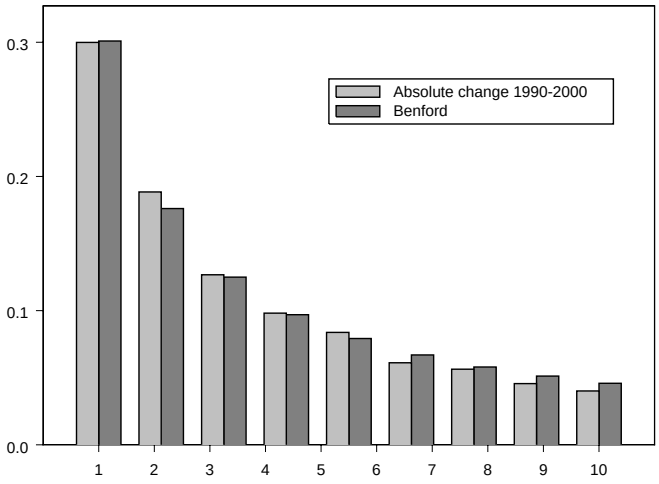


Figure 3.3: Absolute change in the population of the U.S. counties between 1990 and 2000

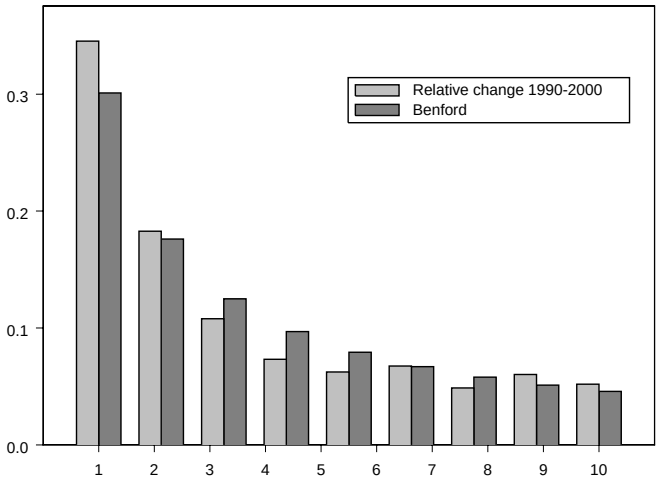


Figure 3.4: Relative change in the population of the U.S. counties between 1990 and 2000

3.1.3 Populations of the U.S. towns

Figure 3.5 shows that populations of the U.S. small towns taken from the 2000 census are very close to Benford's Law, more in a sense than the counties. The reason is certainly that the dataset is much larger, and hence an asymptotical character of Benford's Law takes place. Here of course the Chi-square and the p-value are however penalized a lot by the sample size (25150).

	Chi-square	Distance	p-value
Towns	15.1	$9.1\ 10^{-5}$	6%

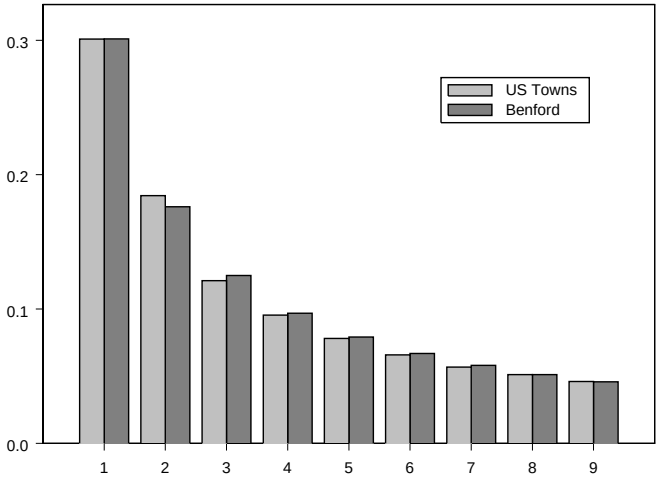


Figure 3.5: Population of 25150 U.S. towns in 2000

3.2 French departements and regions

The populations of French departements and regions (different kind of administrative subdvisions) are here checked for conformity to Benford's Law. The data are taken from INSEE (www.insee.fr).

3.2.1 Departements

A departement is a small administrative subdivision of France (an area of approximatively five thousand square kilometers), and there are 99 of them in total (including overseas territories).

Figure 3.6 and the Chi-square results below lead to the same conclusion: it cannot be supposed that the populations of French departements follow Benford's Law. In fact, the dataset is not large enough to see Benford's Law clearly appear. Such a dataset is subject to too much variance; if for example the populations had been taken each year through the whole century, then Benford's Law would probably have been observed.

Year	Chi-square	Distance	p-value
1990	15.4	$7.9\ 10^{-3}$	25%
1999	10.2	$1.3\ 10^{-2}$	5%

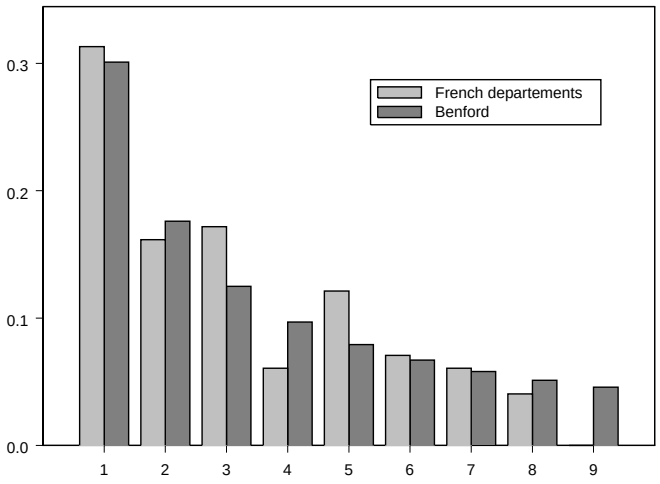


Figure 3.6: Population of the 99 French departements in 1999

3.2.2 Regions

A region is a bigger administrative subdivision of France (typically a region would include 4 departements), and there are 26 of them.

The conclusion is essentially the same as for the departements, since the dataset is even smaller. It has to be noted that for both the departements and the regions the Chi-square test seems to be significant.

Year	Chi-square	Distance	p-value
1990	8.1	$3.7\ 10^{-2}$	42%
1999	7.7	$3.5\ 10^{-2}$	47%

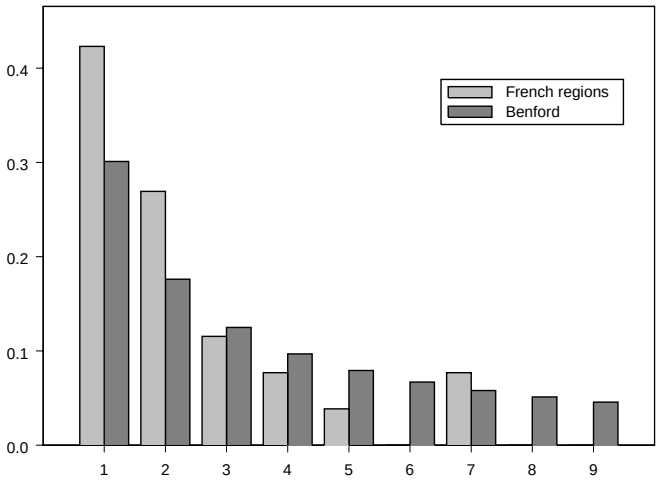


Figure 3.7: Population of the 26 French regions in 1999

3.3 Ensimag address book

In his 1938 experiment, Benford used the street addresses of 342 scientists taken from *American Men of Science*. And the fit to the logarithmic law was very close, one of the best indeed. Since then, it is maybe the dataset that was the most talked about in conferences. Why should such a dataset follow Benford's Law? There is no rigorous answer, the only satisfactory one is perhaps that "there are more short streets than long ones"... Of course it is completely useless, so (justifiably) not much research has been done on that subject.

The author was very skeptical about this (like about Benford's Law in general, by the way), and wanted to show that Benford had 'luck' in his experiment. The street addresses of 515 Ensimag students were thus analyzed... Both the histogram (figure 3.8) and the Chi-square results are quite amazing...

	Chi-square	Distance	p-value
Adresses	6.4	$9.1\ 10^{-4}$	60%

Remark 3.3.1 *The data are precisely the home addresses -in other words the parents' addresses-. These were chosen since for some obscure reason students tend to live together, and hence the dataset of the students' own addresses includes many duplicated entries...*

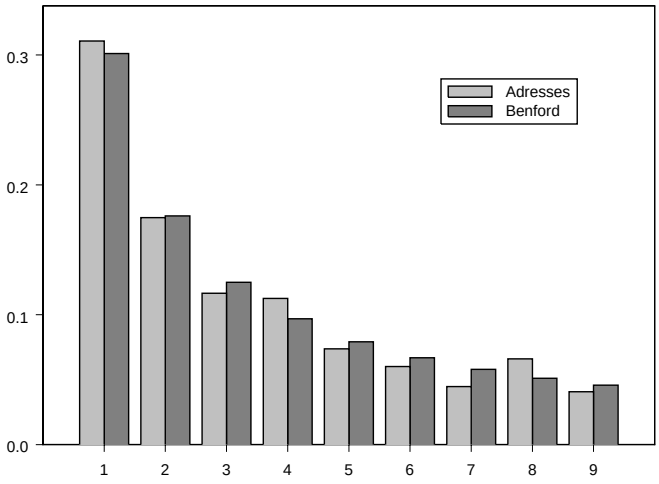


Figure 3.8: Street addresses of 515 Ensimag students in 2000-2001

Chapter 4

Check of invariances and fraud detection

4.1 The dataset used

Some invariances in Benford's Law are so amazing that they deserve to be checked empirically. In this section the dataset used to check those invariances is introduced.

4.1.1 Benford property

The dataset used, which here and throughout will be called D, is a real dataset of 19708 entries, and consists of a mixture of counties and towns taken from the 2000 U.S. Census. The fact that it is a real dataset is not here important in itself; in the current context it is considered as a large Benford dataset with a bit of 'natural' noise (which is better to use than a perfect Benford dataset).

However, the conformity of D to Benford's Law has first to be checked. Figures 4.1 and 4.2 show histograms for the first and second digits of D. The next table confirms what can be observed, i.e. that the first digit has a excellent fit, whereas the second has a medium one.

Digit	Chi-square	Distance	p-value
First	6.1	$3.5\ 10^{-5}$	64%
Second	48.7	$2.5\ 10^{-4}$	0%

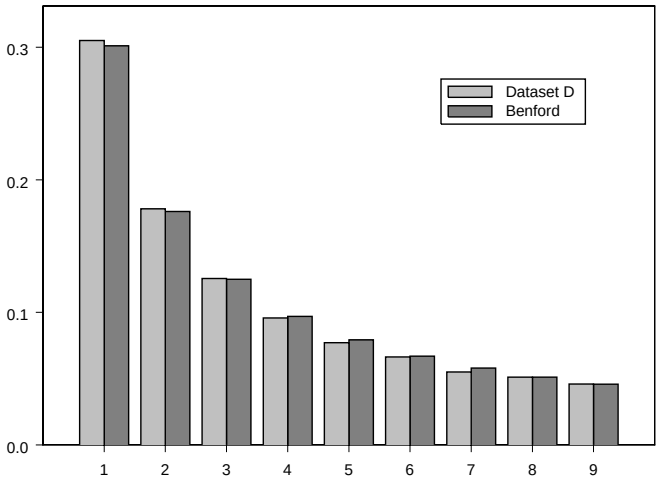


Figure 4.1: First digit in D

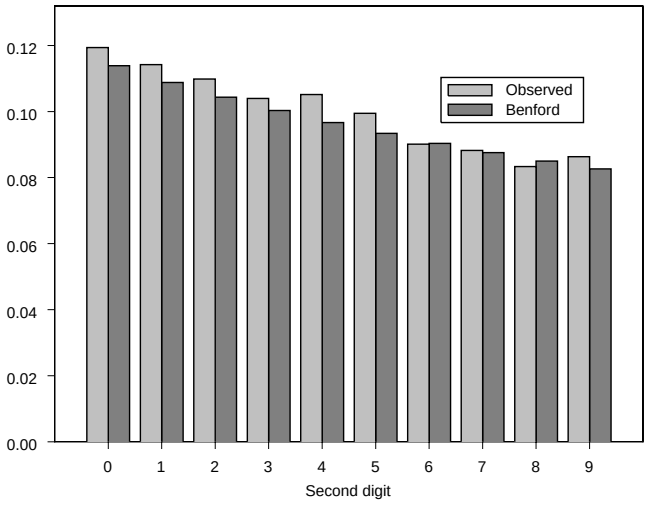


Figure 4.2: Second digit in D

4.1.2 Fraud detection

Benford's Law is used by accountants to detect fraud and/or duplicated data. It is a *partially negative rule*, like many other rules, i.e. if Benford's Law is not satisfied, then it is probable that the dataset was manipulated (so further detailed tests are necessary), but conversely a good Benford dataset could possibly be fraudulent.

Actually this is the case of the dataset D. It was obtained by mistake, mixing some of the populations of towns and counties. What makes it 'fraudulent' is that some of the entries are sums of others (a county is divided into several towns). Hence some of the data in D are heavily duplicated, not in the normal sense but in the sense that they are split into sums. Not only did Benford's Law not help to spot that kind of duplication (which is logical), but also one can certainly say that it made the fit better.

Another experiment was then performed, to see how relevant Benford's Law can be in fraud detection. The original dataset was contaminated by a sample from a normal distribution with the same mean and the same variance. The data were replaced at random in D by the fraudulent ones, so that the size of the dataset was kept constant. This was taken as a model for a clever fraud, but there are numerous other methods one can think about.

Figure 4.3 shows the evolution of the distance from Benford's Law, according to the size of the fraudulent sample (presented here as a percentage of the size of the whole dataset). The range of contamination was chosen to be 0-10%, in order to be realistic.

Several remarks have to be done:

- An exponential curve was found (graphically) to be a very good fit, which means that the distance to Benford's Law increases exponentially as the dataset is more contaminated. This makes this simple distance a possible tool in fraud detection, but all the problem will then be to design a correct warning level.

- The variance in the distance is large, especially for high levels of contamination, so this warning level may not be accurate at all. For instance, a same distance value was approximatively found for a level of 3.5% and for a level of 8.5%. Separating natural noise from actual fraud might thus be really tricky.

- The decimal order of the distance itself somehow contradicts what was said in section 3.1.1, i.e. that 10^{-4} was a medium value for conformity... Here an order of 10^{-4} seems to be quite bad, since it corresponds to a contamination level of 10%. This kind of analysis helps to quantify what should be expected from the distance.

Of course everything that was done here is based on a demographic dataset that nobody (or nearly...) would have the interest to fraud. And the kind of fraud used is only an example, in many cases simple duplications are performed.

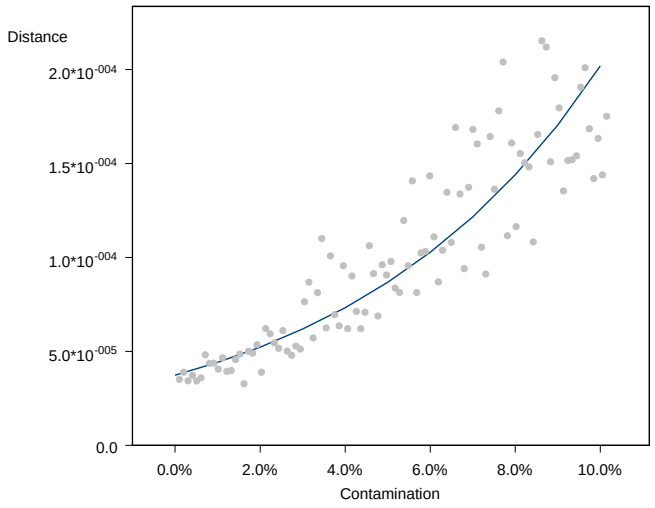


Figure 4.3: Detection of fraudulent data in D

4.2 Scale- and base-invariance

Scale- and base-invariance are some of the most important invariances in Benford's Law. In this section they will be extensively tested on D.

4.2.1 Scale invariance

The dataset D was multiplied by 91 constants regularly spaced between 1 and 10 (hence with a step of 0.1). Figure 4.4 shows the distance to Benford's Law for the first digit of each of the scaled datasets.

What has to be remarked is that the distance varies discontinuously, with an average of $6.4\ 10^{-5}$ and a variance of $5.6\ 10^{-10}$. Scale-invariance is roughly observed, since the decimal order of thr distance doesn't exceed 10^{-4} in the worst case. The scale of the original dataset D (i.e. 1) is luckily one of the best scales.

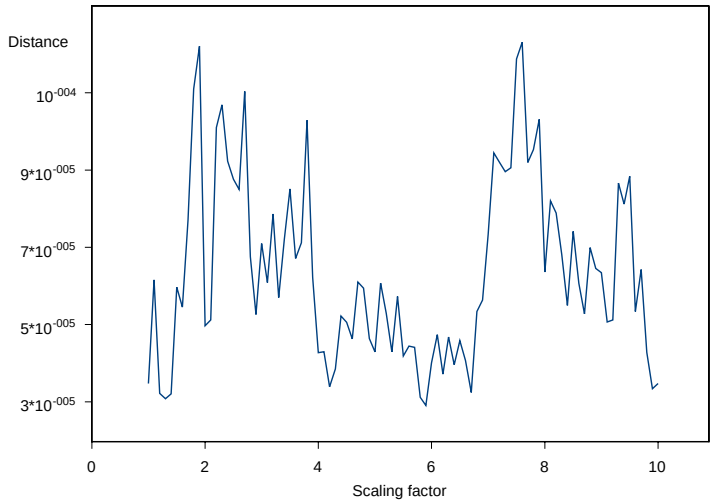


Figure 4.4: Scale-invariance

4.2.2 Base invariance

The dataset D was expressed in 8 different bases, from 3 to 10. Bases greater than 10 were not investigated because it would have required a much more complicated computer routine. Base 2 was not either, as in base 2 the first digit is always 1. The next table shows the Chi-square, the distance, and the p-value for the first digits of each of the converted datasets. Here the Chi-squares and the distances do not have the same meaning according to the base (in base b there are $b-1$ classes for the first digit, so they tend to be naturally lower in the low bases). Taking that into account, the average distance $|f_i - e_i|$ on all the bases and all the classes was found to be of the order of 10^{-3}, so base invariance is verified.

Here the Chi-squares seem not to be significant. In base 3 the excellent fit is probably due to the fact that there are only two classes. The p-values are very bad as usual, but if the distances are considered, overall base-invariance is quite well verified.

Base	Chi-square	Distance	p-value
3	0.1	$2.7\ 10^{-6}$	74 %
4	5.7	$7.1\ 10^{-5}$	6 %
5	13.4	$1.4\ 10^{-4}$	0 %
6	13.9	$1.3\ 10^{-4}$	1 %
7	14.1	$1.1\ 10^{-4}$	1 %
8	20.4	$1.9\ 10^{-4}$	0 %
9	12.9	$8.1\ 10^{-5}$	8 %
10	6.1	$3.5\ 10^{-5}$	64 %

4.3 Inverse and multiplication

Inverse and multiplication invariances are another kind of invariances, which explain the appearance of Benford's Law in long series of calculations.

4.3.1 Inverse

Figure 4.5 shows the histogram for the first digit of the inverse of D, and the next table gathers the Chi-square and distance results. According to the distance, inverse invariance is well verified.

Digit	Chi-square	Distance	p-value
Inverse	12.6	$6.9\ 10^{-5}$	13%

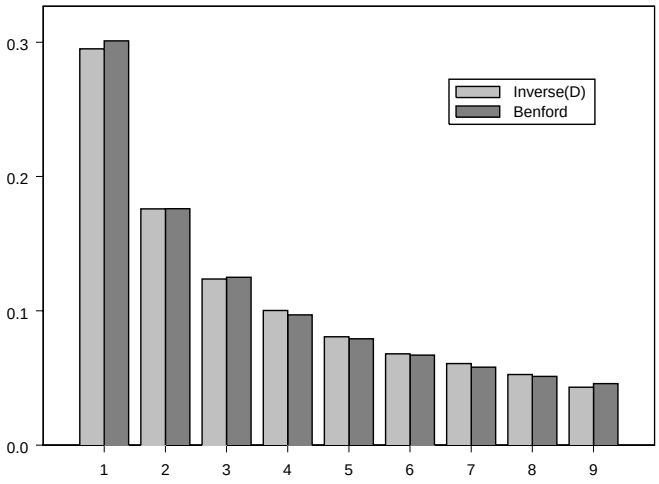


Figure 4.5: Inverse of D

4.3.2 Multiplication

Figures 4.6, 4.7 and 4.8 show histograms for the first digit of the product of D by a sample (of size 19708) from three usual distributions (uniform, normal, and exponential). As proved in section 2.5.2, multiplication invariance (by *any* distribution, which is somehow a very amazing property) is actually observed for the few chosen distributions.

Digit	Chi-square	Distance	p-value
Uniform	10.0	$4.6\ 10^{-5}$	26%
Normal	8.2	$5.7\ 10^{-5}$	41%
Exponential	7.0	$5.8\ 10^{-5}$	54%

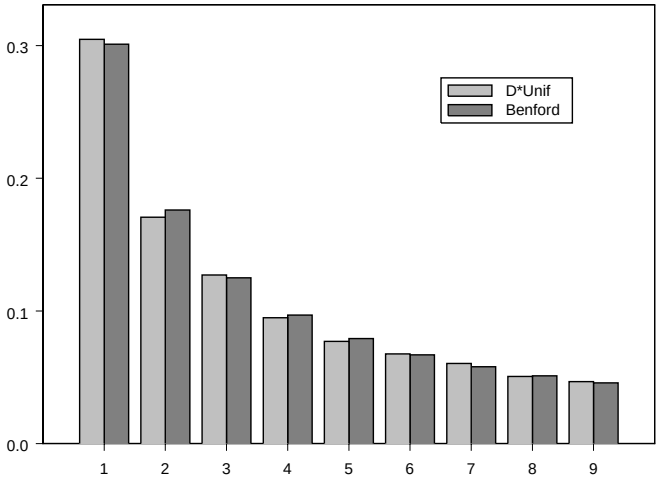


Figure 4.6: Multiplication by a sample from a uniform distribution on $[0, 1]$

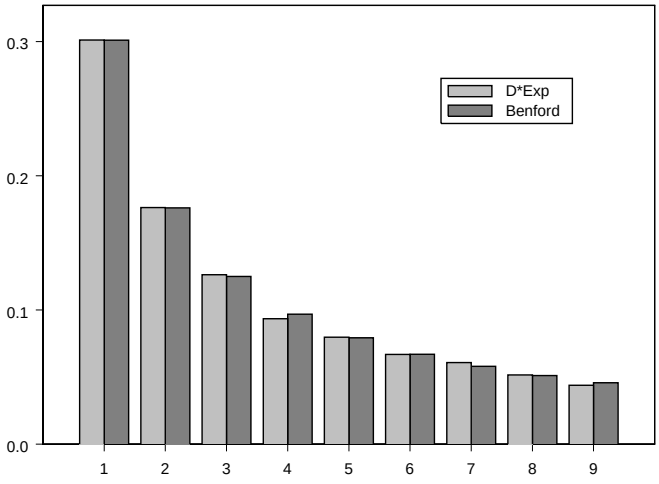


Figure 4.7: Multiplication by a sample from a normal distribution $N(0,1)$

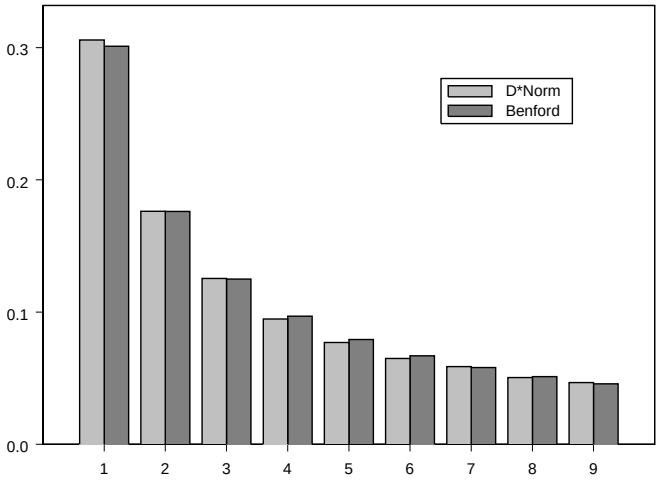


Figure 4.8: Multiplication by a sample from an exponential distribution $Exp(1)$

Conclusion

Benford's Law history, which is quite fascinating, is a good example of how scientists evolve in their approaches. First, the property was investigated on the whole set of integers; then, invariances were found and used to hypothesize Benford's Law; and eventually the idea of randomized distributions was introduced. Benford's Law history is full of controversies, mistakes, and discussions, and there is still a lot to write...

A summary of the mathematical properties described in the present report could be:

1. Real numbers satisfy Benford's Law if their logarithms are uniformly distributed modulo 1 (section 2.1.4);
2. For any 'non-pathological' distribution, the significant digits at the infinite tend to be uniformly distributed (section 2.2);
3. The logarithmic distribution is the only scale-invariant distribution (section 2.3.2);
4. The logarithmic distribution is the only atomless base-invariant distribution (section 2.3.3);
5. In observing samples from multiple distributions, supposing that the average distribution is scale- or base-invariant (and atomless) is sufficient for Benford's Law to appear (section 2.3.4);
6. The logarithmic distribution is the only sum-invariant distribution (section 2.4);
7. The logarithmic distribution is invariant under inversion and multiplication by any distribution (section 2.5);
8. A sequence of products of random variables is very likely to converge to Benford's Law (section 2.5.3);
9. Geometrical sequences are in general logarithmically distributed, along with other sequences (section 2.6).

The underlying question was, among these properties, which can explain the appearance of Benford's Law in so many datasets. A possible partial answer is:

- Property 9 for particular datasets like populations or financial indexes;
- Property 8 for datasets coming from long series of computations;
- Property 5 for datasets consisting of mixtures of other datasets.

In the process of checking the existence of Benford's Law, some datasets were found to follow the logarithmic law very closely. Some of the best fits were the populations of the U.S. towns and the Ensimag address book. The problem of quantifying the goodness-of-fit was raised: Benford's Law only appear clearly enough in large datasets, where the Chi-square statistic is of little use. Another statistic, a simple distance to Benford's Law, was found to be more useful in those cases, but again it seems to lack accuracy and is highly dependent on the dataset. Fraud detection was then somehow criticized, owing to the fact that in such datasets noise and fraudulent data will be hard to distinguish.

In the case of Benford's Law, invariances are so amazing in general that the author found they were needed to be checked. Overall the few chosen were verified, by the distance used. Maybe a further extension could be to check these invariances on other classical distributions and see if some can be applied to them (for instance the property of convergence to the uniform was first thought to be only applicable to digits following a logarithmic distribution, but in fact it was found to be a much more general property...).

There could be two other direct sequels to the project:

- Writing a formal proof for Hill's conjecture 2.2.1,
- Designing an accurate testing method for conformity of large datasets to Benford's Law.

Investigating other distributions for digits could also be interesting. The logarithmic distribution should not be the only useful one... Actually some skeptical reader could say that, given any usual distribution for digits, there would certainly be plenty of confirming datasets to be found. Some distributions of digits are more interesting or more widespread than others, though, and Benford's Law seems to be one of the best in those domains.

Bibliography

[Adhikari & Sarkar] ADHIKARI, A., and SARKAR, B. (1968). Distribution of most significant digit in certain functions whose arguments are random variables. *Indian Journal of Statistics B (Sankhya Ser.)* 30, 47-58.

[Allaart] ALLAART, P.C. (1997). An invariant-sum characterization of Benford's Law. *Journal of Applied Probability* 34, 288-291.

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

[Becker] BECKER, P. (1982). Patterns in listings of failure-rate and MTTF values and listings of other data. *IEEE Transactions on Reliability* R-31, 132-134.

[Buck *et al.*] BUCK, B., MERCHANT, A., and PEREZ, S. (1993). An illustration of Benford's first digit law using alpha decay half lives. *European Journal of Physics* 14, 59-63.

[Burke & Kincanon] BURKE, J., and KINCANON , E. (1991). Benford's Law and physical constants: the ditribution of initial digits. *American Journal of Physics* 59, 952.

[Diaconis] DIACONIS, P. (1977). The distribution of leading digits and uniform distribution mod 1. *Annals of Probability* 5, 72-81.

[Hamming] HAMMING, R. (1970). On the distribution of numbers. *Bell System Technical Journal* 49 1609-1625.

[Hilla] HILL, T.P. (1995). The significant-digit law. *Statistical Science* 10(4), 354-363.

[Hillb] HILL, T.P. (1995). Base-invariance implies benford's Law. *Proceedings of the American Mathematical Society* 123, 887-895.

[Hillc] HILL, T.P. (1995). The significant digit phenomenon. *American Mathematical Monthly* 103, 322-327.

[Knuth] KNUTH, D. (1969). *The art of computer programming*, 219-229. Addison-Wesley, Reading, MA.

[Kuipers & Niederreiter] KUIPERS, L. and NIEDERREITER, H. (1974). *Uniform distributions of sequences*, 1-18. Wiley, New-York.

[Leemis *et al.*] LEEMIS, L.M., SCHMEISER, B.W. and EVANS, D.L. (2000). Survival distributions satisfying Benford's Law. *The American Statistician* 54(3), 236-241.

[Ley] LEY, E. (1996). On the peculiar distribution of the U.S. stock indexes' digits. *The American Statistician* 50(4), 311-313.

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

[Nigrini & Mittermaier] NIGRINI, M.J., and MITTERMAIER, L.J. (1997). The use of Benford's Law as an aid in analytical procedures. *Auditing: A Journal of Practice & Theory* 16(2), 52-67.

[Nigrini] NIGRINI, M.J. (1999). I've got your number. *Journal of Accountancy* May, 79-83.

[Pinkham] PINKHAM, R.S. (1961). On the distribution of first significant digits. *Annals of Mathematical Statistics* 32, 1223-1230

[Raimi] RAIMI, R. (1976). The first digit problem. *American Mathematical Monthly* 83, 521-538.

[Schatte] SCHATTE, P. (1988). On mantissa distributions in computing and Benford's Law. *Journal of Information Processing and Cybernetics* 24, 443-455.

[Varian] VARIAN, H. (1972). Benford's Law. *American Statistician* 23(June), 65-66.

[Weyl] WEYL, H. (1916). Über die Gleichverteilung von Zahlen mod Eins. *Mathematische Annalen* 77, 313-352.