Chapter One

Introduction

Benford's law, also known as the *First-digit* or *Significant-digit law*, is the empirical gem of statistical folklore that in many naturally occurring tables of numerical data, the significant digits are not uniformly distributed as might be expected, but instead follow a particular logarithmic distribution. In its most common formulation, the special case of the first significant (i.e., first non-zero) decimal digit, Benford's law asserts that the leading digit is not equally likely to be any one of the nine possible digits $1, 2, \ldots, 9$, but is 1 more than 30% of the time, and is 9 less than 5% of the time, with the probabilities decreasing monotonically in between; see Figure 1.1. More precisely, the exact law for the first significant digit is

$$\mathsf{Prob}(D_1 = d) = \log_{10}\left(1 + \frac{1}{d}\right) \quad \text{for all } d = 1, 2, \ldots, 9\,; \tag{1.1}$$

here, D_1 denotes the first significant decimal digit, e.g.,

$$\begin{aligned} D_1\big(\sqrt{2}\big) &= D_1(1.414) = 1\,, \\ D_1\big(\pi^{-1}\big) &= D_1(0.3183) = 3\,, \\ D_1\big(e^{\pi}\big) &= D_1(23.14) = 2\,. \end{aligned}$$

Hence, the two smallest digits occur as the first significant digit with a combined probability close to 50 percent, whereas the two largest digits together have a probability of less than 10 percent, since

$$\mathsf{Prob}(D_1 = 1) = \log_{10} 2 = 0.3010\,, \quad \mathsf{Prob}(D_1 = 2) = \log_{10}\frac{3}{2} = 0.1760\,,$$

and

$$\mathsf{Prob}(D_1 = 8) = \log_{10}\frac{9}{8} = 0.05115\,, \quad \mathsf{Prob}(D_1 = 9) = \log_{10}\frac{10}{9} = 0.04575\,.$$

The complete form of Benford's law also specifies the probabilities of occurrence of the second and higher significant digits, and more generally, the *joint* distribution of *all* the significant digits. A general statement of Benford's law that includes the probabilities of all blocks of consecutive initial significant digits is this: For every positive integer m, and for all initial blocks of m significant digits $(d_1, d_2, \dots, d_m)$, where d_1 is in $\{1, 2, \dots, 9\}$, and d_j is in $\{0, 1, \dots, 9\}$ for all $j \geq 2$,

$$\mathsf{Prob}\big(D_1 = d_1, D_2 = d_2, \dots, D_m = d_m\big) = \log_{10}\left(1 + \left(\sum\nolimits_{j=1}^{m} 10^{m-j} d_j\right)^{-1}\right), \tag{1.2}$$

where D_2, D_3, D_4, etc. represent the second, third, fourth, etc. significant decimal digits, e.g.,

$$D_2\big(\sqrt{2}\big) = 4\,, \quad D_3\big(\pi^{-1}\big) = 8\,, \quad D_4\big(e^{\pi}\big) = 4\,.$$

For example, (1.2) yields the probabilities for the individual second significant digits,

$$\mathsf{Prob}\big(D_2 = d_2\big) = \sum\nolimits_{j=1}^{9} \log_{10}\left(1 + \frac{1}{10j + d_2}\right) \quad \text{for all } d_2 = 0, 1, \dots, 9\,, \tag{1.3}$$

which also are not uniformly distributed on all the possible second digit values $0, 1, \dots, 9$, but are strictly decreasing, although they are much closer to uniform than the first digits; see Figure 1.1.

d	0	1	2	3	4	5	6	7	8	9
$\mathsf{Prob}(D_1 = d)$	0	30.10	17.60	12.49	9.69	7.91	6.69	5.79	5.11	4.57
$\mathsf{Prob}(D_2 = d)$	11.96	11.38	10.88	10.43	10.03	9.66	9.33	9.03	8.75	8.49
$\mathsf{Prob}(D_3 = d)$	10.17	10.13	10.09	10.05	10.01	9.97	9.94	9.90	9.86	9.82
$\mathsf{Prob}(D_4 = d)$	10.01	10.01	10.00	10.00	10.00	9.99	9.99	9.99	9.98	9.98

Figure 1.1: Probabilities (in percent) of the first four significant decimal digits, as implied by Benford's law (1.2); note that the first row is simply the first-digit law (1.1).

More generally, (1.2) yields the probabilities for longer blocks of digits as well. For instance, the probability that a number has the same first three significant digits as $\pi = 3.141$ is

$$\mathsf{Prob}\big(D_1 = 3, D_2 = 1, D_3 = 4\big) = \log_{10}\left(1 + \frac{1}{314}\right) = \log_{10}\frac{315}{314} = 0.001380\,.$$

A perhaps surprising corollary of the general form of Benford's law (1.2) is that the significant digits are dependent, and not independent as one might expect

[74]. To see this, note that (1.3) implies that the (unconditional) probability that the second digit equals 1 is

$$\mathsf{Prob}(D_2 = 1) = \sum\nolimits_{j=1}^{9} \log_{10}\left(1 + \frac{1}{10j+1}\right) = \log_{10}\frac{6029312}{4638501} = 0.1138\,,$$

whereas it follows from (1.2) that if the first digit is 1, the (conditional) probability that the second digit also equals 1 is

$$\mathsf{Prob}(D_2 = 1 | D_1 = 1) = \frac{\log_{10} 12 - \log_{10} 11}{\log_{10} 2} = 0.1255\,.$$

Note. Throughout, real numbers such as $\sqrt{2}$ and π are displayed to *four* correct significant decimal digits. Thus an equation like $\sqrt{2} = 1.414$ ought to be read as $1414 \le 1000 \cdot \sqrt{2} < 1415$, and *not* as $\sqrt{2} = \frac{1414}{1000}$. The only exceptions to this rule are probabilities given in percent (as in Figure 1.1), as well as the numbers Δ and Δ_∞, introduced later; all these quantities only attain values between 0 and 100, and are shown to *two* correct digits after the decimal point. Thus, for instance, $\Delta = 0.00$ means $0 \le 100 \cdot \Delta < 1$, but not necessarily $\Delta = 0$.

1.1 HISTORY

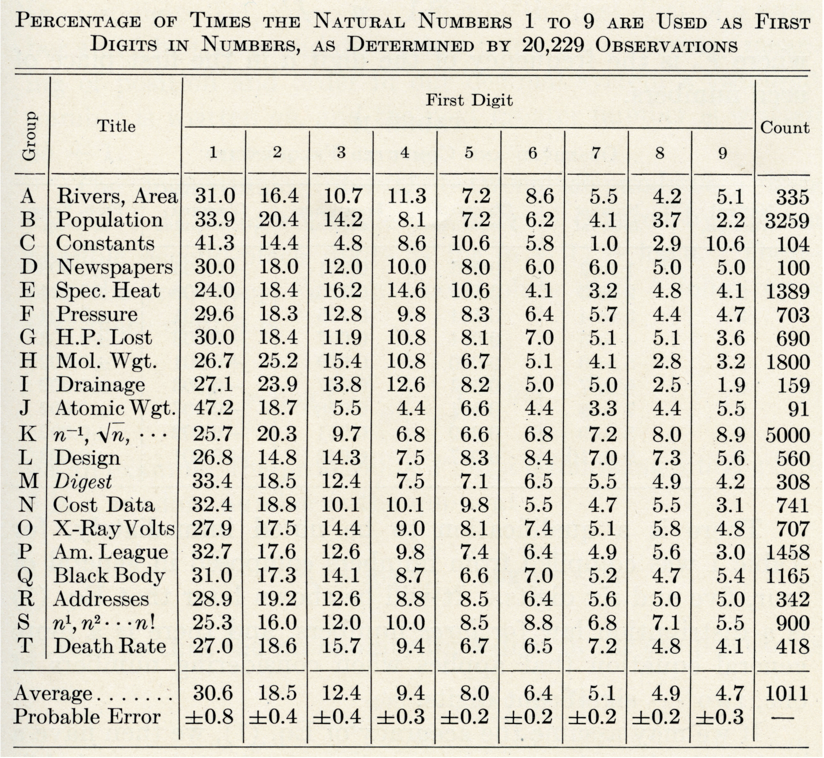
The first known reference to the logarithmic distribution of leading digits dates back to 1881, when the American astronomer Simon Newcomb noticed "how much faster the first pages [of logarithmic tables] wear out than the last ones," and, after several short heuristics, deduced the logarithmic probabilities shown in the first two rows of Figure 1.1 for the first and second digits [111].

Some fifty-seven years later the physicist Frank Benford rediscovered the law [9], and supported it with over 20,000 entries from 20 different tables including such diverse data as catchment areas of 335 rivers, specific heats of 1,389 chemical compounds, American League baseball statistics, and numbers gleaned from front pages of newspapers and *Reader's Digest* articles; see Figure 1.2 (rows A, E, P, D and M, respectively).

Although P. Diaconis and D. Freedman offer convincing evidence that Benford manipulated round-off errors to obtain a better fit to the logarithmic law [47, p. 363], even the unmanipulated data are remarkably close. Benford's article attracted much attention and, Newcomb's article having been overlooked, the law became known as *Benford's law* and many articles on the subject appeared. As R. Raimi observed nearly half a century ago [127, p. 521],

> This particular logarithmic distribution of the first digits, while not universal, is so common and yet so surprising at first glance that it has given rise to a varied literature, among the authors of which are mathematicians, statisticians, economists, engineers, physicists, and amateurs.

The online database [24] now references more than 800 articles on Benford's law, as well as other resources (books, websites, lectures, etc.).



PERCENTAGE OF TIMES THE NATURAL NUMBERS 1 TO 9 ARE USED AS FIRST DIGITS IN NUMBERS, AS DETERMINED BY 20,229 OBSERVATIONS

Group	Title	First Digit									Count
		1	2	3	4	5	6	7	8	9	
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	$n^{-1}, \sqrt{n}, \cdots$	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	*Digest*	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	$n^1, n^2 \cdots n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	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		±0.8	±0.4	±0.4	±0.3	±0.2	±0.2	±0.2	±0.2	±0.3	—

Figure 1.2: Benford's original data from [9]; reprinted courtesy of the American Philosophical Society.

1.2 EMPIRICAL EVIDENCE

Many tables of numerical data, of course, do *not* follow Benford's law in any sense. Telephone numbers in a given region typically begin with the same few digits, and never begin with a 1; lottery numbers in all common lotteries are distributed uniformly, not logarithmically; and tables of heights of human adults, whether given in feet or meters, clearly do not begin with a 1 about 30% of the time. Even "neutral" mathematical data such as square-root tables of integers do not follow Benford's law, as Benford himself discovered (see row K in Figure 1.2 above), nor do the prime numbers, as will be seen in later chapters.

On the other hand, since Benford's popularization of the law, an abundance of additional empirical evidence has appeared. In physics, for example, D. Knuth [90] and J. Burke and E. Kincanon [31] observed that of the most commonly used physical constants (e.g., the speed of light and the force of gravity listed on the inside cover of an introductory physics textbook), about 30% have leading significant digit 1; P. Becker [8] observed that the decimal parts of failure (hazard) rates often have a logarithmic distribution; and R. Buck et al., in studying the values of the 477 radioactive half-lives of unhindered alpha decays that were accumulated throughout the past century, and that vary over many orders of magnitude, found that the frequency of occurrence of the first digits of both measured and calculated values of the half-lives is in "good agreement" with Benford's law [29]. In scientific calculations, A. Feldstein and P. Turner called the assumption of logarithmically distributed mantissas "widely used and well established" [57, p. 241]; R. Hamming labeled the appearance of the logarithmic distribution in floating-point numbers "well-known" [70, p. 1609]; and Knuth observed that "repeated calculations with real numbers will nearly always tend to yield better and better approximations to a logarithmic distribution" [90, p. 262].

Additional empirical evidence of Benford's law continues to appear. M. Nigrini observed that the digital frequencies of certain entries in Internal Revenue Service files are an extremely good fit to Benford's law (see [113] and Figure 1.3); E. Ley found that "the series of one-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" [98]; and Z. Shengmin and W. Wenchao found that "Benford's law reasonably holds for the two main Chinese stock indices" [148]. In the field of biology, E. Costas et al. observed that in a certain cyanobacterium, "the distribution of the number of cells per colony satisfies Benford's law" [39, p. 341]; S. Docampo et al. reported that "gross data sets of daily pollen counts from three aerobiological stations (located in European cities with different features regarding vegetation and climatology) fit Benford's law" [49, p. 275]; and J. Friar et al. found that "the Benford distribution produces excellent fits" to certain basic genome data [60, p. 1].

Figure 1.3 compares the probabilities of occurrence of first digits predicted by (1.1) to the distributions of first digits in four datasets: the combined data reported by Benford in 1938 (second-to-last row in Figure 1.2); the populations of the 3,143 counties in the United States in the 2010 census [102]; all numbers appearing on the World Wide Web as estimated using a Google search experiment [97]; and over 90,000 entries for Interest Received in U.S. tax returns from the IRS Individual Tax Model Files [113]. To instill in the reader a *quantitative* perception of closeness to, or deviation from, the first-digit law (1.1), for every distribution of the first significant decimal digit shown in this book, the number

$$\Delta = 100 \cdot \max_{d=1}^{9} \left| \mathsf{Prob}(D_1 = d) - \log_{10}\left(1 + \frac{1}{d}\right) \right|$$

will also be displayed. Note that Δ is simply the maximum difference, in percent, between the probabilities of the first significant digits of the given distribution and the Benford probabilities in (1.1). Thus, for example, $\Delta = 0$ indicates exact conformance to (1.1), and $\Delta = 12.08$ indicates that the probability of some digit $d \in \{1, 2, \ldots, 9\}$ differs from $\log_{10}(1 + d^{-1})$ by 12.08%, and the probability of no other digit differs by more than this.

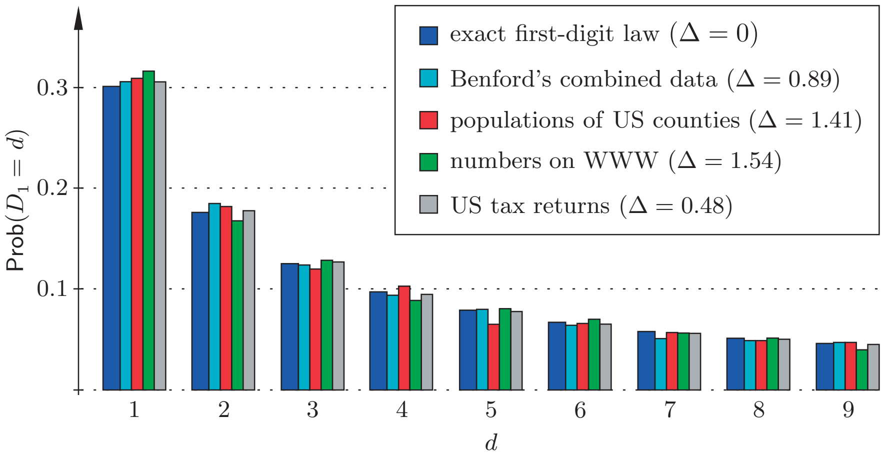


Figure 1.3: Comparisons of four datasets to Benford's law (1.1).

All these statistics aside, the authors also highly recommend that the justifiably skeptical reader perform a simple experiment, such as randomly selecting numerical data from front pages of several local newspapers, or from "a Farmer's Almanack" as Knuth suggests [90], or running a Google search similar to the Dartmouth classroom project described in [97].

1.3 EARLY EXPLANATIONS

Since the empirical significant-digit law (1.1) or (1.2) does not specify a well-defined statistical experiment or sample space, most early attempts to explain the appearance of Benford's law argued that it is "merely the result of our way of writing numbers" [67] or "a built-in characteristic of our number system" [159]. The idea was to first show that the set of real numbers satisfies (1.1) or (1.2), and then suggest that this explains the empirical statistical evidence. A common starting point has been to try to establish (1.1) for the positive integers, beginning with the prototypical set

$$\{D_1 = 1\} = \{1, 10, 11, \ldots, 18, 19, 100, 101, \ldots, 198, 199, 1000, 1001, \ldots\},$$

the set of positive integers with first significant digit 1. The source of difficulty and much of the fascination of the first-digit problem is that the set $\{D_1 = 1\}$ does not have a *natural density* among the integers, that is, the proportion of integers in the set $\{D_1 = 1\}$ up to N, i.e., the ratio

$$\frac{\#\{1 \le n \le N : D_1(n) = 1\}}{N}, \tag{1.4}$$

does not have a limit as N goes to infinity, unlike the sets of even integers or primes, say, which have natural densities $\frac{1}{2}$ and 0, respectively. It is easy to see that the empirical density (1.4) of $\{D_1 = 1\}$ oscillates repeatedly between $\frac{1}{9}$ and $\frac{5}{9}$, and thus it is theoretically possible to assign any number between $\frac{1}{9}$ and $\frac{5}{9}$ as the "probability" of this set. Similarly, the empirical density of $\{D_1 = 9\}$ forever oscillates between $\frac{1}{81}$ and $\frac{1}{9}$; see Figure 1.4.

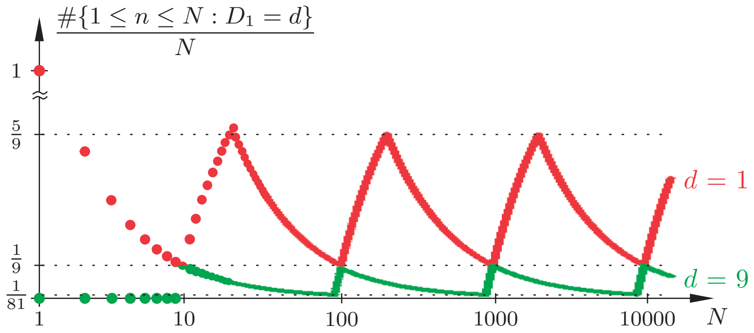


Figure 1.4: The sets $\{D_1 = 1\}$ and $\{D_1 = 9\}$ do not have a natural density (and neither does $\{D_1 = d\}$ for any $d = 2, 3, \ldots, 8$).

Many partial attempts to put Benford's law on a solid logical basis have been made, beginning with Newcomb's own heuristics, and continuing through the decades with various urn model arguments and mathematical proofs; Raimi [127] has an excellent review of these. But as the eminent logician, mathematician, and philosopher C. S. Peirce once observed, "in no other branch of mathematics is it so easy for experts to blunder as in probability theory" [63, p. 273], and the arguments surrounding Benford's law certainly bear that out. Even W. Feller's classic and hugely influential text [58] contains a critical flaw that apparently went unnoticed for half a century. Specifically, the claim by Feller and subsequent authors that "regularity and large spread implies Benford's Law" is fallacious for any reasonable definitions of regularity and spread (measure of dispersion) [21].

1.4 MATHEMATICAL FRAMEWORK

A crucial part of (1.1), of course, is an appropriate interpretation of Prob. In practice, this can take several forms. For sequences of real numbers $(x_1, x_2, \ldots)$, Prob usually refers to the limiting proportion (or relative frequency) of elements in the sequence for which an event such as $\{D_1 = 1\}$ occurs. Equivalently, fix a positive integer N and calculate the probability that the first digit is 1 in an experiment where one of the elements $x_1, x_2, \ldots, x_N$ is selected at random (each with probability $1/N$); if this probability has a limit as N goes to infinity, then the limiting probability is designated $\text{Prob}(D_1 = 1)$. Implicit in this usage of

Prob is the assumption that all limiting proportions of interest actually exist. Similarly, for real-valued functions $f : [0, +\infty) \to \mathbb{R}$, fix a positive real number T, choose a number τ at random uniformly between 0 and T, and calculate the probability that $f(\tau)$ has first significant digit 1. If this probability has a limit, as $T \to +\infty$, then $\mathsf{Prob}(D_1 = 1)$ is that limiting probability.

For a random variable or probability distribution, on the other hand, Prob simply denotes the underlying probability of the given event. Thus, if X is a random variable, then $\mathsf{Prob}\,(D_1(X) = 1)$ is the probability that the first significant digit of X is 1. Finite datasets of real numbers can also be dealt with this way, with Prob being the empirical distribution of the dataset.

One of the main themes of this book is the robustness of Benford's law. In the context of sequences of numbers, for example, iterations of linear maps typically follow Benford's law exactly; Figure 1.5 illustrates the convergence of first-digit probabilities for the Fibonacci sequence $(1, 1, 2, 3, 5, 8, 13, \ldots)$. As will be seen in Chapter 6, not only do iterations of most linear functions follow Benford's law exactly, but iterations of most functions *close* to linear also follow Benford's law *exactly.* Similarly, as will be seen in Chapter 8, powers and products of very general classes of random variables approach Benford's law in the limit; Figure 1.6 illustrates this starting with $U(0, 1)$, the standard random variable uniformly distributed between 0 and 1. Similarly, if random samples from different randomly-selected probability distributions are combined, the resulting meta-sample also typically converges to Benford's law; Figure 1.7 illustrates this by comparing two of Benford's original empirical datasets with the combination of all his data.

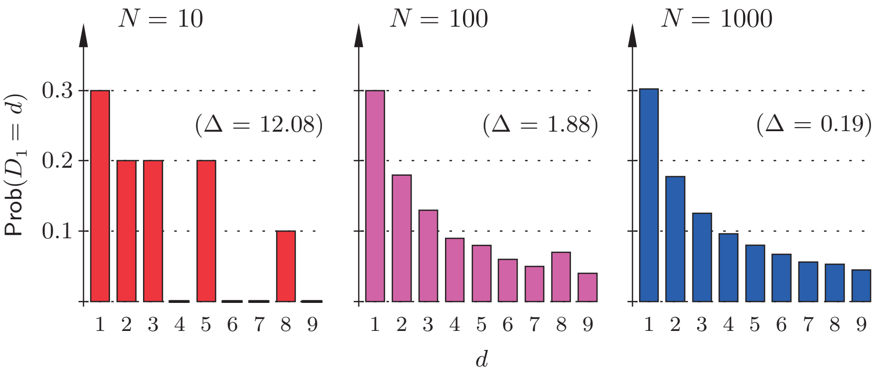


Figure 1.5: Probabilities that a number chosen uniformly from among the first N Fibonacci numbers has first significant digit d.

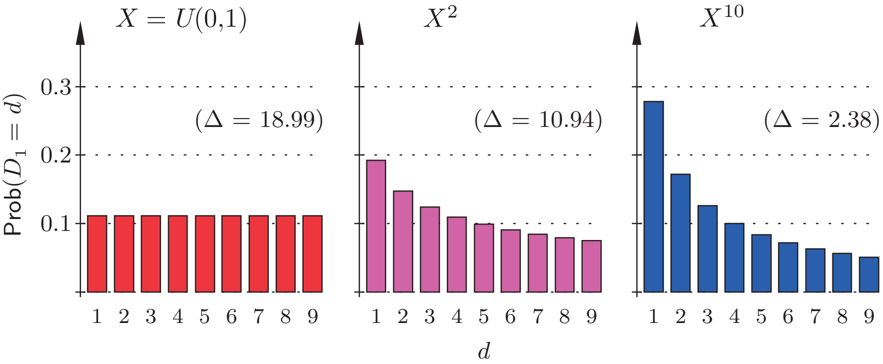


Figure 1.6: First-digit probabilities of powers of a $U(0,1)$ random variable X.

Non-decimal bases

Throughout this book, attention will normally be restricted to decimal (i.e., base-10) significant digits, and when results for more general bases are employed, that will be made explicit. From now on, therefore, $\log x$ will always denote the logarithm base 10 of x, while $\ln x$ is the natural logarithm of x. For convenience, the convention $\log 0 := \ln 0 := 0$ is adopted. Nearly all the results in this book that are stated only with respect to base 10 carry over easily to arbitrary integer bases $b \geq 2$, and the interested reader may find some pertinent details in [15]. In particular, the general form of (1.2) with respect to any such base b is

$$\mathsf{Prob}\left(D_1^{(b)} = d_1, D_2^{(b)} = d_2, \ldots, D_m^{(b)} = d_m\right) = \log_b\left(1+\left(\sum\nolimits_{j=1}^{m} b^{m-j}d_j\right)^{-1}\right), \tag{1.5}$$

where $\log_b$ denotes the base-b logarithm, and $D_1^{(b)}$, $D_2^{(b)}$, $D_3^{(b)}$, etc. are the first, second, third, etc. significant digits base b, respectively; so in (1.5), d_1 is an integer in $\{1, 2, \ldots, b-1\}$, and for $j \geq 2$, d_j is an integer in $\{0, 1, \ldots, b-1\}$. Note that in the case $m = 1$ and $b = 2$, (1.5) reduces to $\mathsf{Prob}\left(D_1^{(2)} = 1\right) = 1$, which is trivially true because the first significant digit base 2 of every non-zero number is 1.

This book is organized as follows. Chapter 2 contains formal definitions, examples, and graphs of significant digits and the significand (mantissa) function, and also of the probability spaces needed to formulate Benford's law precisely, including the crucial natural domain of "events," the so-called significand σ-algebra. Chapter 3 defines Benford sequences, functions, and random variables, with examples of each. Chapters 4 and 5 contain four of the main mathematical characterizations of Benford's law, with proofs and examples. Chapters 6 and 7 study Benford's law in the context of deterministic processes, including both one- and multi-dimensional discrete-time dynamical systems and algorithms as

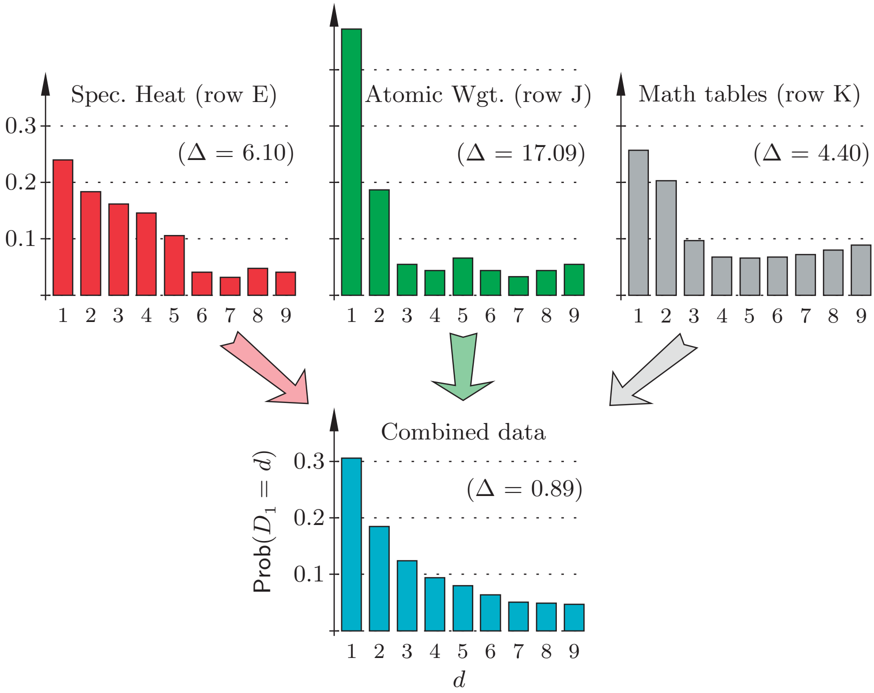


Figure 1.7: Empirical first-digit probabilities in Benford's original data; see Figure 1.2.

well as continuous-time processes generated by differential equations. Chapter 8 addresses Benford's law for random variables and stochastic processes, including products of random variables, mixtures of distributions, and random maps. Chapter 9 offers a glimpse of the complementary theory of Benford's law in the non-traditional context of finitely additive probability theory, and Chapter 10 provides a brief overview of the many applications of Benford's law that continue to appear in a wide range of disciplines.

The mathematical detail in this book is on several levels: The basic explanations, and many of the figures and comments, are intended for a general scientific audience; the formal statements of definitions and theorems are accessible to an undergraduate mathematics student; and the proofs, some of which contain basics of measure and ergodic theory, are accessible to a mathematics graduate student (or a diligent undergraduate).