Introduction

It has been frequently observed that in many tables of physical constants and statistical data, the leading significant digit is not uniformly distributed among the digits $\{1, 2, \ldots, 9\}$ as might be expected; rather the lower digits appear much more frequently than the higher ones. Perhaps even more surprising, an exact distribution for this non-uniformity of the leading digits has been generally asserted. In 1881 Simon Newcomb [9] stated that "The law of probability of the occurrence of numbers is such that the mantissae of their logarithms are equally probable," and concluded that

(1) $\text{Prob (first significant digit} = d) = \log_{10}(1 + d^{-1}), \qquad d = 1, 2, \ldots, 9.$

(For example, (1) predicts that the leading digit is 1 with probability about .301, and at the other extreme, is 9 with probability .046.)

Although Newcomb offered no statistical evidence for (1), its rediscovery by the physicist Benford [2] some fifty-seven years later was supported by empirical evidence based on frequencies of significant digits from twenty different tables including such diverse data as surface areas of 335 rivers, specific heat of thousands of chemical compounds, and square-root tables. The union of his tables comes surprisingly close to the frequencies predicted in (1), and, Newcomb's earlier paper having been overlooked, those frequencies came to be known as *Benford's Law*, or the *First Digit Law*. In fact, Benford's data not only came surprisingly close, it came *suspiciously* close to the predicted frequencies; Diaconis and Freedman [5, p. 363] offer convincing evidence that Benford manipulated the round-off errors to obtain an even better fit. But even the unmanipulated data seems a remarkably good fit, and the "law" has become widely accepted.

Classical Explanations

Since Benford, numerous "mathematicians, statisticians, economists, engineers, physicists and amateurs" [11, p. 521] have attempted to explain the probabilities appearing in (1) based on a variety of hypotheses. The classical explanations include: the usual number-theoretic (or Cesaro) method for assigning densities to the sets in question; continuous analogs of the Cesaro method based on integration techniques; various probabilistic

1

urn-schemes; demonstrations based on assumptions of continuity and scale-invariance (see below); and statistical descriptive arguments. For an excellent review of these ideas, the reader is referred to Raimi [11]. (A more recent explanation of Schatte [12] gives Benford's Law as a corollary to an unproved ([12, p. 452]) "hypothesis that after a sufficiently long computation in floating-point arithmetic, the occurring mantissas have a nearly logarithmic distribution.")

All of these previous explanations suffer from two substantial shortcomings. First, the previous methods for prescribing frequencies for such sets as "first significant digit = 1" are *not unique.* Such a set does not have a natural density, unlike the set of even numbers, say, which has density 1/2 among the integers and density 0 among the real numbers, and in general there are many ways of assigning a number to the set "first significant digit = d" which are consistent with natural density. The explanations mentioned above simply single out particular summation or integration techniques that yield the "correct" Benford frequencies.

The second shortcoming is that, terminology notwithstanding, the past frequency-assigning functions leading to (1) are *not probabilities*, at least not in the classical sense. The standard mathematical definition of probability is a [0,1]-valued function P on a domain of sets (called a sigma algebra) closed under complements and countable unions, which assigns 1 to the whole set and assigns measure $\sum_{n=1}^{\infty} P(A_n)$ to the set $\bigcup_{n=1}^{\infty} A_n$ if the $\{A_n\}$ are disjoint. But the methods above necessarily fail to satisfy these conditions, as will, for example, any reasonable notion of density on the natural numbers which assigns density 0 to singletons, for then $P(\mathbb{N}) = \sum^{\infty} P(\{n\}) = 0 \neq 1$. (This is exactly the same reason for the foundational difficulty in making rigorous sense of "pick an integer at random"; e.g., see De Finetti [4] pages 86, 98–99). For the integer-based models of Benford's Law, this difficulty seems insurmountable, and for the above-mentioned real-number models either a precise domain for the probability in (1) was not specified by Newcomb *et al.*, or when specified was simply not the appropriate collection $\mathcal{A}$.

The Proper Probability Domain

The first step toward making rigorous sense of the First-Digit Law (1) is to identify an appropriate domain for the probability. A typical set in the desired collection $\mathcal{A}$ of subsets of $\mathbb{R}^+$ is the set of positive reals whose first significant digit (base 10) is 1, namely,

$$\{D_1 = 1\} := \bigcup_{n=-\infty}^{\infty} [1,2) \cdot 10^n.$$

This set (along with its analogs from the second, and general n^{th}-digit laws, also known to Newcomb and Benford) suggests the following natural domain $\mathcal{A}$ for a general significant-digit law.

Definition. $\mathcal{A}$ is the smallest collection of subsets of the positive reals which contains all sets of the form $\bigcup_{n=-\infty}^{\infty} (a,b) \cdot 10^n$, and which is closed under complements and countable unions.

The following properties of $\mathcal{A}$ are easy to check:

every non-empty set in $\mathcal{A}$ is infinite, with accumulation points at 0 and at $+\infty$;

$\mathcal{A}$ is closed under scalar multiplication, i.e., $a > 0$ and $S \in \mathcal{A} \Rightarrow aS \in \mathcal{A}$;

$\mathcal{A}$ is self-similar, in the sense that if $S \in \mathcal{A}$ and $k \in \mathbb{Z}$ then $10^k S = S$.

For each $i = 1, 2, \ldots$, let $D_i : \mathbb{R}^+ \to \{0, 1, \ldots, 9\}$ be the $i-th$ significant digit function, for example, $D_1(\pi) = 3, D_2(\pi) = 1 = D_2(10\pi)$. It may easily be shown [8] that

$$D_i^{-1}(\{d\}) \in \mathcal{A} \text{ for all } i \text{ and } d,$$

and in fact, $\mathcal{A}$ is the smallest such collection (closed under complements and countable unions) for which this is true. (In measure-theoretic terms, $\mathcal{A}$ is the sigma-algebra generated by $D_1, D_2, \ldots$) This shows that $\mathcal{A}$ is precisely the correct domain for a general significant-digit probability law.

The General Significant-Digit Law

General Significant-Digit Law [8]. For all $k \in N$, all $d_1 \in \{1, 2, \ldots, 9\}$ and all $d_j \in \{0, 1, 2, \ldots, 9\}$, $j = 2, \ldots, k$,

$$(2) \qquad P\Big(\bigcap_{i=1}^{k}\{D_i = d_i\}\Big) = \log_{10}\Big[1 + \Big(\sum_{i=1}^{k} d_i \cdot 10^{k-i}\Big)^{-1}\Big].$$

Observe that this *joint* significant-digit law (2) includes the First-Digit Law (1) as a special case, as well as the other marginal significant-digit laws.

Example. $\qquad P((D_1, D_2, D_3) = (3, 1, 4)) = \log_{10}(1 + \frac{1}{314}) \cong .0014.$

A perhaps surprising corollary of (2) is that

the significant digits are dependent

and not independent as one might expect. For example, from (2) it follows that the (unconditional) probability that the second digit is 2 is $\cong .109$, but the (conditional) probability the second digit is 2, *given* that the first digit is 1, is $\cong .115$. Similarly, the hundredth significant-digit is also dependent on the first few significant digits, although the dependency decreases as distance between the digits increases. It also follows easily from (2) that the distribution of the ith significant digit approaches the uniform distribution (where each digit $\{0, 1, \ldots, 9\}$ occurs with frequency $\frac{1}{10}$) exponentially fast as $i \to \infty$.

What simple hypotheses lead to the General Significant-Digit Law (2)?

Scale and Base-Invariance

One set of hypotheses which has been popular in the past is the notion of *scale-invariance*, which corresponds to the following idea. If the first digits obey some fixed universal distributional law, then this law should be independent of the units chosen (e.g., English or metric systems). However, as Knuth pointed out (cf. Raimi [11]), *there is no scale-invariant probability measure on the Borel subsets of* $\mathbb{R}^+$, since then the measure of the set $(0, 1)$ must be the same as the measure of *every* interval $(0, b)$, which by countable additivity must be 0.

4

The problem is simply that the Borel sets (the smallest sigma-algebra containing all open intervals) are not the appropriate domain for a significant-digit probability law; using $\mathcal{A}$ instead resolves this problem.

On $\mathcal{A}$, it is easily shown [8] (using the fact that the orbit of every point under irrational rotation on the circle is dense) that if P is scale-invariant, i.e., if $P(bS) = P(S)$ for all $b > 0$ and all $S \in \mathcal{A}$, then P satisfies (2). That is, on the correct domain $\mathcal{A}$,
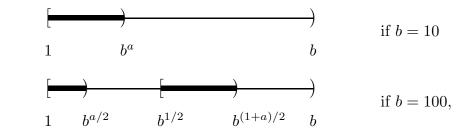
scale-invariance implies Benford's Law.

One possible drawback to the scale-invariance hypothesis is the special role played by the constant 1. In most tables of physical constants, the constant 1 simply does not appear, since the underlying law (say, in $f = ma$) does not necessitate definition of a constant (as opposed to $e = mC^2$). If a "complete" table of physical constants included the constant 1, perhaps that special constant would occur with strictly positive frequency. But this would preclude scale-invariance, since then $0 < P(\{1\}) = P(\{2\}) = \ldots$, contradicting the additivity of a probability.

As an alternative hypothesis, suppose that any universal significant-digit law were *base-invariant*; i.e., carried over to bases other than 10. (As pointed out in [11], all the classical arguments supporting (1) and (2) carry over *mutatis mutandis* to other bases such as 2, or 7 or 100.)

To motivate a formal definition of base-invariance, consider the set of positive numbers S with first significant digit (base 10) less than 5. Using the decimal notation D_1 as above, and letting $D_1^{(100)}$ denote the first significant digit base 100, it is easily seen that

$$S = \{1 \leq D_1 < 5\} = \{1 \leq D_1^{(100)} < 5\} \cup \{10 \leq D_1^{(100)} < 50\},$$

which says that graphically (as a subset of $[1, b)$), the *same* set S is

and is

5

(where $a = \log_{10} 5$). Hence if a probability P on $\mathcal{A}$ is "base-invariant," the measures of these two S-representing subsets of $[1, b)$ should be the same, i.e.,

$$P([1, b^a)) = P([1, b^{\frac{1}{2}})) + P([b^{\frac{1}{2}}, b^{\frac{1+a}{2}})),$$

and similarly for higher power bases b^n. This suggests the following definition.

Definition. [8] *P is base-invariant on $\mathcal{A}$ if*

$$P([1, 10^a]) = \sum_{k=0}^{n-1} P[10^{\frac{k}{n}}, 10^{\frac{k+a}{n}}) \text{ for all } n \in \mathbb{N} \text{ and all } a \in (0, 1).$$

Letting P_L be the logarithmic probability defined in (2) and P_0 be the degenerate probability which assigns mass 1 to the constant 1 (or formally, to the set $\bigcup_{n=-\infty}^{\infty} \{10^n\}$ in $\mathcal{A}$), it now follows [8] using a somewhat deeper result from ergodic theory concerning invariant measures on the circle, that

$$P \text{ is base-invariant} \iff P = qP_0 + (1-q)P_L \quad \text{for some } q \in [0, 1].$$

Corollaries are:

   the logarithmic distribution (2) is the unique continuous base-invariant distribution

and

   scale-invariance implies base-invariance.

(Observe that base-invariance does *not* imply scale-invariance, since P_0 is base but not scale-invariant.) Thus, if there is a universal significant-digit law and it is base-invariant, then the special constant 1 occurs with possibly positive probability q, and otherwise (with probability $1 - q$) the digits satisfy the logarithmic distribution (2).

Applications

Computer design and analysis of roundoff errors

   Hamming [6] has given applications of Benford's Law to the problem of placing the decimal (binary) point in the number system of a computer in order to minimize the number

of normalization shifts after the computation of a product, to the problem of estimation of the representation error of numbers in base 2 and base 16, and to the problem of roundoff error propagation. Schatte [12] similarly concludes that the choice of a binary-power base $b = 2^r$ can be guided by the hypothesis of logarithmic distribution (cf. Benford's Law) of mantissa errors; for example, he argues that base $b = 2^3$ is optimal with respect to storage use.

Statistical Tests for "Naturalness"

Varian [13] has proposed using Benford's Law as a test of "reasonableness" for data, by checking forecasts of a mathematical model as to goodness of fit to Benford's Law. He used this idea to check specific models for economic production and for forecasts of acres of land in various use, and Becker [1] used Benford's Law to check lists of failure rates to detect systematic errors. The underlying idea in these applications is that if "real life data" obeys Benford's Law, then so should good mathematical models.

Making Money in Numbers Games

In the Massachussets Numbers Game [cf. 3], players first bet on a four-digit number of their choice, next a single four-digit number is generated randomly by an umpire, and then all players with the winning number share the (tax-reduced) pot equally. In such a situation it is obviously advantageous to identify numbers which few people choose, since all numbers are equally likely to be winners and the expected payoff for an unpopular number is thus higher than that for a number which many people have chosen. Now *if* people choose numbers from their experience, and *if* the numbers in their experience obey Benford's Law, then it makes sense to pick numbers *inversely* to Benford's Law, i.e., numbers starting with 9 or 8. Of Chernoff's [3] 33 statistically-obtained numbers in his "first system" (numbers with predicted normalized payoffs exceeding 1.0) for playing the Massachussets Numbers Game, 16 had first significant digit 8 or 9, and only 1 has first significant digit 1 or 2. (Additional evidence that numbers "randomly" generated by people tend to start with low digits is found in Hill [7].) Since Chernoff also concluded that the public learns quickly, this suggests using inverse-Benford as an initial strategy when a new numbers game is initiated, and then quitting play soon thereafter.

Outfoxing the Internal Revenue Service

In his Ph.D. thesis, Nigrini [10] has suggested that the IRS use Benford's Law as a test for detecting falsification of data by a taxpayer at the time of filing his return. Nigrini's hypothesis is that true data gives a rough approximation to Benford's Law, whereas a Benford-ignorant cheater tends to concoct numbers according to some other distribution, say uniform via a standard random number generator, or more likely, a subconscious personal favorite generated mentally. Nigrini proposes that the IRS simply check for goodness-of-fit against Benford, and then audit the worst fits. This suggests that a "creative" and Benford-wise taxpayer should modify (or generate) his fabricated data according to a Benford-like distribution.

Acknowlegement

References

1. 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.

2. Benford, F. (1938) The law of anomalous numbers. *Proc. Amer. Phil. Soc.* **78**, 551-72.

3. Chernoff, H. (1981) How to beat the Massachusetts Numbers Game. *Math. Intel.* **3**, 166-172.

4. De Finetti, B. (1972) *Probability, Induction and Statistics.* Wiley, New York.

5. Diaconis, P. and Freedman, D (1979) On rounding percentages. *J. Amer. Stat. Assoc.*, 359-64.

6. Hamming, R. (1976) On the distribution of numbers. *Bell Syst. Tech. J.* **49**, 1609-25.

7. Hill, T. (1988) Random-number guessing and the first digit phenomenon. *Psychological Reports* **62**, 967-71.

8. Hill, T. (1993) Base-invariance implies Benford's Law, to appear in *Proc. Amer. Math. Soc.*

9. Newcomb, S. (1881) Note on the frequency of use of the different digits in natural numbers. *Amer. J. Math.* **4**, 39-40.

10. Nigrini, M. (1992) The detection of income evasion through an analysis of digital distributions. Ph.D. Thesis, Department of Accounting, University of Cincinnati.

11. Raimi, R. (1976) The first digit problem. *Amer. Math. Monthly* **83**, 521-38.

12. Schatte, P. (1988) On mantissa distributions in computing and Benford's Law. *J. Inf. Process. Cybern.* EIK **24**, 443-455.

13. Varian, H. (1972) Benford's Law. *Amer. Statistician* **26**, 65-66.

The Significant-Digit Phenomenon[1]

by
Theodore P. Hill
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332