THE FIRST-DIGIT PHENOMENON

T. P. Hill

A century-old observation concerning an unexpected pattern in the first digits of many tables of numerical data has recently been discovered to also hold for the stock market, census statistics, and accounting figures. New mathematical insights establish this empirical law as part of modern probability theory, and recent applications include testing of mathematical models, design of computers, and detection of fraud in tax returns.

In 1881, the astronomer/mathematician Simon Newcomb published a 2-page article in the *American Journal of Mathematics* describing his observation that books of logarithms in the library were dirtier in the beginning and progressively cleaner throughout. From this he inferred that researchers using the logarithm tables (fellow astronomers and mathematicians, as well as biologists, sociologists and other scholars) were looking up numbers starting with 1 more often than numbers starting with 2, and numbers with first digit 2 more often than 3, and so on. After a short heuristic argument, Newcomb concluded that the probability that a number has first significant digit (i.e., first non-zero digit) d is

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

In particular, his conjecture was that the first digit is 1 about 30% of the time, and is 9 only about 4.6% of the time (Figure 1). That the digits are not equally likely comes as somewhat of a surprise, but to claim an exact law describing their distribution is indeed striking.

Newcomb's article went unnoticed, and 57 years later General Electric physicist Frank Benford, apparently unaware of Newcomb's paper, made exactly the same observation about logarithm books and also concluded the same logarithm law. But then Benford *tested* his conjecture with an "effort to collect data from as many as fields as possible

and to include a wide variety of types ... the range of subjects studied and tabulated was as wide as time and energy permitted." Evidence indicates that Benford spent several years gathering data, and the table he published in 1938 in the *Proceedings of the American Philosophical Society* was based on 20,229 observations from such diverse data sets as areas of rivers, American League baseball statistics, atomic weights of elements, and numbers appearing in *Reader's Digest* articles. His resulting table of first significant digits (the number to the left of the decimal point in "scientific notation;" Figure 2) fit the logarithm law exceedingly well.

Unlike Newcomb's article, Benford's paper drew a great deal of attention (partly due to the good fortune of having it published adjacent to a soon-to-be-famous physics paper), and Newcomb's contribution having become completely forgotten, the logarithm probability law came to be known as Benford's Law.

(There is also a general significant-digit law which includes not only the first digits, but also the second (which may be 0) and all higher significant digits, and even the *joint* distribution of the digits. The general law says, for example, that the probability that the first three significant digits are 3, 1, and 4, in that order, is

$$\log_{10}\left(1+\frac{1}{314}\right)\cong 0.0014$$

and similarly for other significant-digit patterns. From this general law it follows that the second significant digits, although decreasing in relative frequency through the digits as do the first-digits, are much more uniformly distributed than the first digits, and the third than the second, etc. The general law also implies that the significant digits are *not independent* as might be expected, but that instead knowledge of one digit affects the likelihood of another. For example, an easy calculation shows that the (unconditional) probability that the second digit is 2 is $\cong 0.109$, but the (conditional) probability that the second digit is 2 *given* that the first digit is 1, is $\cong 0.115$.)

Of course, many tables of numerical data do *not* follow this logarithmic distribution – lists of telephone numbers in a given region usually begin with the same few digits, and even "neutral" data such as square root tables are not a good fit. However, a surprisingly diverse collection of empirical data does obey the logarithm law for significant digits, and

since Benford's popularization of the law, an abundance of empirical evidence has appeared: tables of physical constants, numbers appearing in newspaper front pages, accounting data and scientific calculations.

The assumption of logarithmically-distributed significant-digits (i.e., floating point numbers) in scientific calculations is "widely used and well-established," and Stanford computer scientist Donald Knuth's classic text *The Art of Computer Programmer* devotes a section to the log law. More recently, analyst Eduardo Ley of *Resources for the Future* in Washington, DC has found that stock market figures (*Dow-Jones Industrial Average* index, and *Standard and Poor's* index) fit Benford's law closely, and accounting professor Mark Nigrini, whose PhD thesis was based on applications of Benford's law, discovered that in the 1990 U.S. Census, the populations of the three thousand counties in the U.S. are also a very good fit to Benford's law (Figure 1).

The skeptical reader is encouraged to perform a simple experiment such as listing all the numbers occurring on the front pages of several local newspapers, or randomly selecting data from "a Farmer's Alamack," as Knuth suggests.

In the 60 years since Benford's article appeared there have been numerous attempts by mathematicians, physicists, statisticians and amateurs to "prove" Benford's law, but there have been two main stumbling blocks. The first is very simple – some data sets satisfy the law, and some do not, and there never was a clear definition of a general statistical experiment which would predict which tables would, and which would not. Instead, researchers endeavored to prove that the log law "is a built-in characteristic of our number system;" that is, to prove that the set of *all* numbers satisfies the log law, and then to suggest that this somehow explains the frequent empirical evidence. Attempts at proofs were based on various mathematical averaging and integration techniques, as well as probabilistic "drawing balls from urns" schemes.

One popular hypothesis in this context has been that of assuming *scale invariance*, which corresponds to the intuitively attractive idea that if there is indeed any universal law for significant digits, then it certainly should be independent of the units used (e.g., metric or English). It was observed empirically that tables which fit the log law closely also fit it closely if converted (by scale multiplication) to other units, or converted to reciprocal

units. For example, if stock prices closely follow Benford's law (as Ley found they do), then conversion from dollars per stock to pesos (or yen) per stock should not alter the first-digit frequencies much, even though the first digits of individual stock prices will often change radically (Figure 3a). Similarly, converting from dollars per stock to stocks per dollar in Benford tables will also retain nearly the same digital frequencies, whereas in stock tables not originally close to Benford's law (such as uniformly distributed prices, Figure 3b), changing currencies or converting to reciprocals will often dramatically alter the digital frequencies.

Although there was some limited success in showing that Benford's law is the *only* set of digital frequencies which remains fixed under scale changes, the second stumbling block in making mathematical sense of the law is that none of the proofs were rigorous as far as the current theory of probability is concerned. Although both Newcomb and Benford phrased the question as a probabilistic one ("what is the probability that the first significant digit of a number is d?"), modern probability theory requires the intuitive "countable additivity" axiom that if a positive integer (not digit) is picked at random and $p(1)$ is the probability that the number 1 is picked, $p(23)$ the probability 23 is picked, etc., then

$$p(1) + p(2) + p(3) + \cdots = 1,$$

whereas all proofs prior to 1995 failed to satisfy this basic axiom.

One possible drawback to a hypothesis of scale-invariance in tables of universal constants is the special role played by the constant 1. For example, consider the two physical laws $f = ma$ and $e = mC^2$. Both laws involve universal constants, but the force equation constant 1 is not recorded in tables, whereas the speed of light constant C is. If a "complete" list of universal physical constants also included the 1's, it is quite possible that this special constant 1 will occur with strictly positive probability p. But if the table is scale-invariant, then multiplying by a conversion factor of 2 would mean that the constant 2 would also have this same positive probability p, and similarly for all other integers. However this would violate the countable additivity axiom, since then $p(1) + p(2) + \cdots = p + p + p + \cdots =$ infinity, not 1.

Instead, suppose it is assumed that any reasonable universal significant-digit law

should be independent of the base, that is, should be equally valid when expressed in base 10, base 100, binary base 2, or any other base. (In fact, all of the previous arguments supporting Benford's law carry over *mutatis mutandis* to other bases, as many of the authors state.) In investigating this new base-invariance hypothesis, it was discovered that looking at equivalent significant-digit *sets* of numbers, rather than individual numbers themselves, eliminates the previous problems of countable additivity, and allows a formal rigorous proof that the log law is the only probability distribution which is scale-invariant, and the only one which is base-invariant (excluding the constant 1).

(The formal base-invariance theorem in fact states that the only probability distributions on significant digits which are base-invariant are those in which the special constant 1 occurs with possibly positive probability, and the rest of the time the distribution is the log law. The generality of this result implies that *any* other property which is found to imply Benford's law is necessarily base-invariant, and hence a corollary of this theorem.)

These two new results were clean mathematically, but they hardly helped explain the appearance of Benford's law empirically. What do 1990 census statistics have in common with 1880 users of logarithm tables, numerical data from front pages of newspapers of the 1930's collected by Benford, or computer calculations observed by Knuth in the 1960's? What do these tables have in common, and why should they be logarithmic, or equivalently, scale or base invariant?

Many tables are not of this form, including even Benford's individual tables (as he noted), but as University of Rochester mathematician Ralph Raimi pointed out, "what came closest of all, however, was the union of all his tables." Combine the molecular weight tables with baseball statistics and the areas of rivers, and *then* there is a good fit to Benford's law.

Instead of thinking of some universal table of all possible constants (Raimi's "stock of tabular data in the world's libraries" or Knuth's "some imagined set of real numbers"), what seems more natural is to think of data as coming from *many different distributions*, as in Benford's study, in collecting numerical data from newspapers, or in listing stock prices.

Using this idea, modern mathematical probability theory, and the recent scale and

base-invariance proofs, it is not difficult to derive the following new statistical form of the significant-digit law. If distributions are selected at random (in any "unbiased" way) and random samples are taken from each of these distributions, then the significant-digit frequencies of the combined sample will converge to Benford's distribution (Figure 4), even though the individual distributions selected may not closely follow the law.

For example, suppose you are collecting data from a newspaper, and the first article concerns lottery numbers (which are generally uniformly distributed), the second article concerns a particular population with a standard bell curve distribution, and the third is an update of latest calculations of atomic weights. None of these distributions has significant digit frequencies close to Benford's law, but their *average* does (Figure 5), and sampling randomly from each will yield digital frequencies close to Benford's law.

It is of course crucial that the sampling be neutral or unbiased as to scale or base. For example, sampling volumes of soft drink products from various manufacturers in Europe will surely *not* be scale invariant, since those product volumes are closely related to liters, and conversion to other units such as gallons will yield a totally different range of first-digit frequencies. On the other hand, if samples of various species of mammals are selected at random and their volumes determined, it seems much more likely that the resulting data will be unrelated to liters or gallons or other choices of units, and hence that the significant digits will closely follow Benford's law.

One of the points of the new "random samples from random distributions" theorem is that there are *many* natural sampling procedures which lead to the same log distribution, whereas previous arguments were based on the assumption that tables following the log law were all representative of the same mystical underlying set of all constants. Thus the random-sample theorem helps explain how the logarithm-table digital frequencies observed a century ago by Newcomb, and modern tax, census, and stock data all lead to the same log distribution. The new theorem also helps *predict* the appearance of the significant-digit phenomenon in many different empirical contexts (including your morning newspaper), and thus helps justify some of the recent applications of Benford's law.

One of the applications of the significant-digit law has been to testing of mathematical models (Figure 6). Suppose that a new model is proposed to predict future stock indices,

census data, or computer usage. If current data follows Benford's law closely or if a hypothesis of unbiased random samples from random distributions seems reasonable, then the predicted data should also follow Benford's law closely (or else perhaps the model should be replaced by one which does). Such a "Benford-in, Benford-out" test is at best only a double-check on reasonableness, since the law says nothing about the raw data itself; e.g., Benford's law does not distinguish between the numbers 20 and 200,000 – both have first significant digit 2, and all other digits 0.

Another application of Benford's law that has been recently studied is to the design of computers. If computer users of tomorrow are likely to be performing calculations taken from many (unbiased random) distributions, as Knuth and other computer scientists claim is the case today, then their floating-point calculations will be based on data which closely follows Benford's law. In particular, the numbers they will be computing will not be uniformly distributed over the floating point numbers, but will rather follow the log distribution. If this is indeed the case, then it is possible to build computers whose designs capitalize on knowing the distribution of numbers they will be manipulating. If 9's are much less frequent than 1's (or the analog for whatever base the computer is using – recall the principle of base-invariance of Benford's law), then it should be possible to construct computers which use that information to minimize storage space, or to maximize rate of output printed, for example.

The underlying idea is simple – think instead of a cash register drawer. If the frequency of transactions involving the various denominations of bills is known, then the drawer may be specially designed to take advantage of that fact by using bins of different sizes, or which are located in a particular arrangement (such as typewriter and computer keyboards). In fact, German mathematician Peter Schatte has determined that based on an assumption of Benford input, the computer design that minimizes expected storage space (among all computers with binary-power base) is base 8, and other researchers are currently exploring use of logarithmic computers to speed calculations.

A current development in the field of accounting is the application of Benford's law to detect fraud or fabrication of data in financial documents. Nigrini has amassed extensive empirical evidence of the occurrence of Benford's law in many areas of accounting and

demographic data, and has come to the conclusion that in a wide variety of accounting situations, the significant digit frequencies of true data conform very closely to Benford's law (Figure 7).

When people fabricate data, on the other hand, either for fraudulent purposes or just to "fill in the blanks," the concocted data rarely conforms to Benford's law. That people cannot act truly randomly, even in situations where it is to their advantage to do so, is a well-established fact in psychology.

One of my own favorite examples of this from my field of probability is this. The first day of class in an introductory semester of probability theory, I asked the students to do the following homework assignment that evening. If their mother's maiden name begins with A through L, they are to flip a coin 200 times and record the results. The rest of the class is to *fake* a sequence of 200 heads and tails. The next day I collect the results, and separate the fakers' data from the others with 95% accuracy by the following rule. A sequence of 200 truly random tosses of a fair coin contains a run of 6 heads or 6 tails with very high probability (the exact calculation is quite involved), yet the average person trying to fake a random sequence very rarely writes such long runs.

Nigrini's PhD thesis in accounting was based on an analogous idea using Benford's law. Assuming that true accounting data follows Benford's law fairly closely (as his research indicated it does), then substantial deviations from that law suggest possible fraud or fabrication of data. Nigrini has designed several goodness-of-fit tests to measure conformity with Benford's law, and *The Wall Street Journal* (July 10, 1995) reported that the District Attorney's office in Brooklyn, New York was able to detect fraud in seven New York companies using Nigrini's tests (Figure 7). From the evidence to date, it appears that both fraudulent and random-guess data tend to have far too few numbers beginning with 1, and far too many beginning with 6. Based on these preliminary successes, Nigrini has been asked to consult with the internal revenue services of several countries, and is currently helping install Benford goodness-of-fit tests in major accounting fraud-detection computer packages.

At the time of Professor Raimi's article in *Scientific American* on Benford's law over a quarter century ago, the significant-digit phenomenon was thought to be merely a math-

ematical curiosity without real-life application, and without a satisfactory mathematical explanation. He wrote, "Thus all the explanations of [Benford's law] so far given seem to lack something of finality," and concluded "the answer remains obscure."

Although perhaps the final chapter on the significant-digit phenomenon has not been written, today the answer is much less obscure, is firmly couched in the modern mathematical theory of probability, and is seeing important applications to society.

Bibliography

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

T. Hill (1995) Base-invariance implies Benford's law, *Proceedings of the American Mathematical Society* **123**, 887–895.

T. Hill (1996) A statistical derivation of the significant-digit law, *Statistical Science* **10**, 354–363.

E. Ley (1996) On the peculiar distribution of the U.S. stock indices digits, *American Statistician*, to appear.

M. Nigrini (1996) A taxpayer compliance application of Benford's law, *Journal of the American Taxation Association* **18**, 72–91.

R. Raimi (1969) The peculiar distribution of first digits, *Scientific American*, December, 109–119.