

American Scientist

The Magazine of Sigma Xi, The Scientific Research Society

FEATURE ARTICLE

The First Digit Phenomenon

A century-old observation about an unexpected pattern in many numerical tables applies to the stock market, census statistics and accounting data

T. P. Hill

If asked whether certain digits in numbers collected randomly from, for example, the front pages of newspapers or from stock-market prices should occur more often than others, most people would think not. Nonetheless, in 1881, the astronomer and mathematician Simon Newcomb published a two-page article in the *American Journal of Mathematics* reporting an example of exactly that phenomenon. Newcomb described his observation that books of logarithms in the library were quite dirty at the beginning and progressively cleaner throughout. From this he inferred that fellow scientists using the logarithm tables were looking up numbers starting with 1 more often than numbers starting with 2, 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 a particular first significant digit (that is, first nonzero digit) d can be calculated as follows:

Prob (first significant digit = d) = $\log_{10} (1 + 1/d)$, $d = 1, 2, \dots, 9$

In particular, his conjecture was that the first digit is 1 about 30 percent of the time and 9 only about 4.6 percent of the time (see *Figure 2*). That the digits are not equally likely to appear comes as something 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 fields as possible and to include a wide variety of types....[T]he 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 and the first digit in standard notation) fit the logarithm law exceedingly well.



Unlike Newcomb's article, Benford's drew a great deal of attention—partly as a result of the good fortune of being adjacent to a soon-to-be-famous physics paper—and, Newcomb's contribution having been completely forgotten, the logarithm law of probability came to be known as Benford's law.

Although Benford's law has received wide recognition and seen many applications in the last half of the 20th century, a mathematically rigorous proof long proved elusive. In this article I shall describe in general terms the mathematical foundations of the first-digit phenomenon and go on to review some of its recent applications, including the detection of accounting fraud.

Many Good Fits

Although many tables of data follow the logarithmic distribution of first digits, there are likewise many examples that do not. Lists of telephone numbers in a given region, for example, usually begin with the same few digits, and even "neutral" data such as square root tables are not a good fit. Nonetheless, 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 on newspaper front pages, accounting data and scientific calculations.

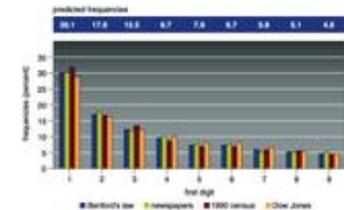
The assumption of logarithmically distributed significant digits in scientific calculations is widely used and well-established, and Stanford computer scientist Donald Knuth's classic 1968 text, *The Art of Computer Programming*, devotes a section to the log law. More recently, in 1996, analyst Eduardo Ley of Resources for the Future in Washington, D.C., found that stock market figures (the Dow Jones Industrial Average index and Standard and Poor's index) fit Benford's law closely. And accounting professor Mark Nigrini, whose Ph.D. thesis was based on applications of Benford's law, discovered that in the 1990 U.S. Census, the populations of the 3,000 counties in the U.S. are also a very good fit to Benford's law. The skeptical reader is encouraged to perform a simple experiment such as listing all the numbers appearing on the front pages of several local newspapers or randomly selecting data from a Farmer's Almanac, as Knuth suggests.

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

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, the third than the second, and so on. For example, the fifth significant digit is still more likely to be a low digit than a high one, but here the differences among digital probabilities are very small—that is, all are close to one-tenth. For sixth significant digits, the probabilities are even closer to being uniformly equal to one-tenth. The general law also implies that the significant digits are not independent, as might be expected, but instead that 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 approximately 0.109, but the (conditional) probability that the second digit is 2 given that the first digit is 1 is approximately 0.115.

Scratching for Proof

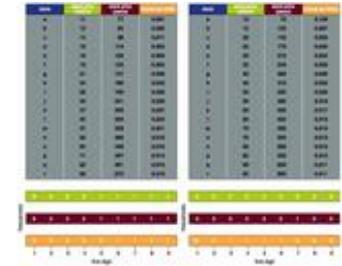
In the 60 years since Benford's article appeared there have been numerous attempts by mathematicians, physicists and amateurs to "prove" Benford's law, but there have been two main stumbling blocks. The first is very simple—some data sets satisfy the law and some do not, and there never was a clear definition of a general statistical experiment that would predict which would and which would not. Instead, mathematicians endeavored to prove that the log law is a built-



$$\log_{10} = (1 + 1/314) \cong 0.0014$$

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 (for example, metric or English). It was observed empirically that tables that fit the log law closely also fit it closely if converted (by scale multiplication) to other units or 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 change radically (see *Figure 3, left*). 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 3, right*), 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 that remains fixed under scale changes, the second stumbling block in making mathematical sense of the law is that none of the proofs was 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 that 23 is picked, and so on, then

$$p(1) + p(2) + p(3) + \dots = 1$$

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 constants also included the 1s, 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. This would violate the countable additivity axiom, since $p(1) + p(2) + \dots = \text{infinity}$, not 1.

Instead, suppose it is assumed that any reasonable universal significant-digit law should be independent of base—that is, it 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 have stated.) 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 that is scale-invariant and the only one that is base-invariant (excluding the constant 1). (The formal base-invariance theorem in fact states that the only probability distributions on significant digits that are base-invariant are those in which the special constant 1 occurs with possible positive probability, and the rest of the time the distribution is the log law [Hill 1995]. The generality of this result implies that any other property that 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 the front pages of newspapers of the 1930s collected by Benford or computer calculations observed by Knuth in the 1960s? Furthermore, why should they be logarithmic or, equivalently, base-invariant?

As already noted, many tables are not of this form, including even Benford's individuals tables, but as University of Rochester mathematician Ralph Raimi pointed out, "what came closest of all, however, was the union of all his tables." Combine molecular-weight tables with baseball statistics and the areas of rivers, and then there is a good fit with 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 (Hill 1996). 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, 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 the latest calculations of atomic weights. None of these distributions has significant-digit frequencies close to Benford's law, but their average does, and sampling randomly from all three will yield digital frequencies close to Benford's law.

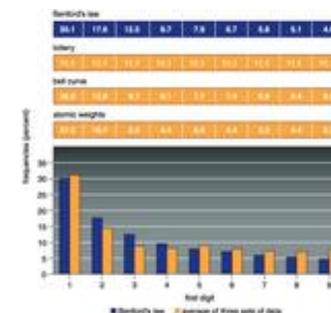
One of the points of the new *random samples from random distributions* theorem is that there are many natural sampling procedures that lead to the same log distribution, whereas previous arguments were based on the assumption that the tables following the log law were all representative of the same 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.

Putting Benford's Law to Work

One of the applications of the significant-digit law has been to the testing of mathematical models (*see Figure 4*). Suppose that a new model is proposed to predict future stock indices, census data or computer usage. If current data follow 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 that 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 themselves. (For example, Benford's law does not distinguish between the numbers 20 and 200,000—both have 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 that closely follow 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 9s are much less frequent than 1s (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 the rate of output printed, for example.

The underlying idea is simple—think instead of a cash register. 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 that are of different sizes or that are located in a particular arrangement (such as typewriter and computer keyboards). In fact, German mathematician Peter Schatte has determined that based on the assumption of Benford input, the computer design that minimizes expected storage space (among all computers with a binary-power base) is base 8, and others are currently exploring the use of logarithmic computers to speed calculations.

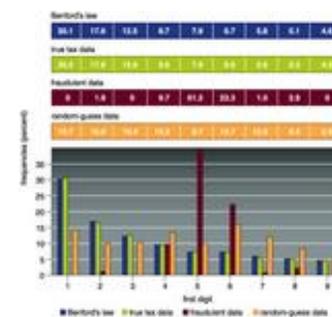
A current development in the field of accounting is the application of Benford's law to detect fraud or fabrication of data in financial documents. Nigrini has amassed extensive empirical evidence of the occurrence of Benford's law in many areas of accounting and demographic data and has come to the conclusion that in a wide variety of accounting situations, the significant-digit frequencies of true data conform very closely to Benford's law (see Figure 5). When people fabricate data, on the other hand, either for fraudulent purposes or just to "fill in the blanks," the concocted data rarely conform 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 own field of probability is in the classroom. The first day of class in an introductory semester of probability theory, I ask 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 percent accuracy using the following rule. A sequence of 200 truly random coin tosses of a fair coin contains a run of six heads or six 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 Ph.D. thesis in accounting was based on an analogous idea using Benford's law. Assuming that true accounting data follow Benford's law fairly closely (as his research indicated that they do), then substantial deviations from that law suggest possible fraud or fabrication of data. He has designed several goodness-of-fit tests to measure conformity with Benford's law, and the *Wall Street Journal* 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. 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 (see Figure 5). 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 Raimi's article on Benford's law in *Scientific American* over a quarter-century ago, the significant-digit phenomenon was thought to be merely a mathematical curiosity without real-life applications, 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 that "the answer remains obscure."

Although the final chapter on the significant-digit phenomenon may not have 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.



You can find this online at <http://www.americanscientist.org/issues/num2/1998/4/the-first-digit-phenomenon/1>

© Sigma Xi, The Scientific Research Society