Benford's Law

Dr. Theodore Hill asks his mathematics students at the Georgia Institute of Technology to go home and either flip a coin 200 times and record the results, or merely pretend to flip a coin and fake 200 results. The following day, he runs his eye over the homework data, and to the students' amazement, he easily identifies nearly all those who faked their tosses.

"The truth is," he said in an interview, "most people don't know the real odds of such an exercise, so they can't fake data convincingly."

There is more to this than a classroom trick.

Dr. Hill is one of a growing number of statisticians, accountants, and mathematicians who are convinced that an astonishing mathematical theorem known as *Benford's Law* is a powerful and relatively simple tool for pointing suspicion at frauds, embezzlers, tax evaders, sloppy accountants, and even computer bugs. The income tax agencies of several nations and and a score of large companies and accounting businesses are using detection software based on Benford's Law to identify faked data.

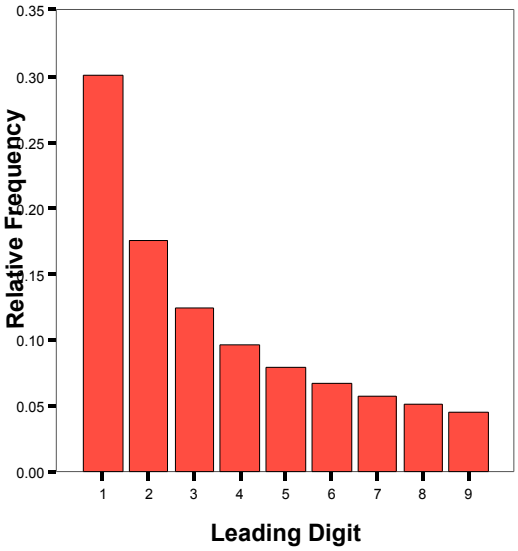
Here's a challenge – go look up some numbers. Just about any variety of naturally-occurring numbers will do. Try the lengths of some of the world's rivers, the cost of gas bills in Moldova, the population sizes in Peruvian provinces, or even the figures in your own tax return. Then, when you have a sample of numbers, look at their first digits (ignoring any leading zeros). Count how many numbers begin with a 1, how many begin with a 2, how many begin with a 3, and so on – what do you find?

You might expect that there would be roughly the same frequency of numbers beginning with each digit; that the proportion of numbers beginning with any given digit would be 1/9. However, in many cases, you'd be wrong.

Surprisingly, for many kinds of data, the distribution of first digits is highly skewed, with 1 being the most common leading digit and 9 the least common. In fact, a precise mathematical relationship seems to hold:

The expected proportion of numbers beginning with the leading digit n **is** $\log\left(1+\frac{1}{n}\right)$ **or** $\log\left(\frac{n+1}{n}\right)$.

This relationship, shown in the graph below and known as Benford's Law, is becoming more and more useful as we understand it better. But how was it discovered and why on earth should it be true?



Leading Digit	1	2	3	4	5	6	7	8	9
$\log\left(\frac{n+1}{n}\right)$	.301	.176	.125	.097	.079	.067	.058	.051	.046

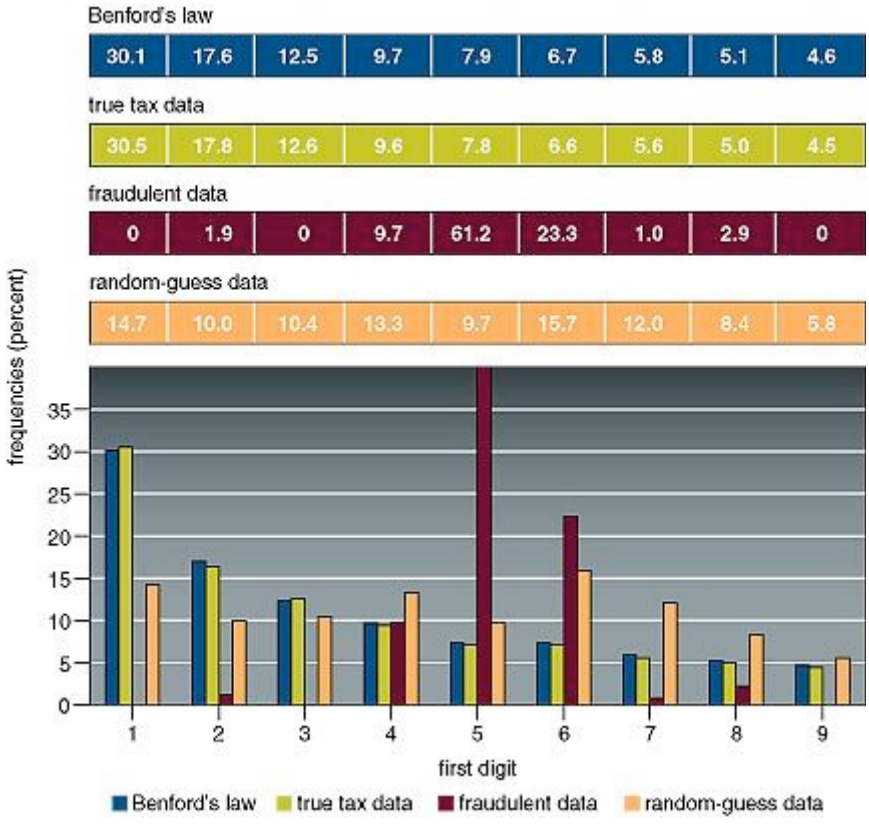
Simon Newcomb, a mathematician and astonomer, was the first person to notice this phenomenon in 1881. While using a book of logarithms for some calculations, Newcomb noticed that pages towards the front of the book were more worn than the pages in the back. Apparently, people did more calculations using numbers that began with lower digits than with higher digits. Newcomb found a formula that matched his observations fairly well. He claimed, without explanation, that the percentage of numbers that start with the digit *d* follows: Log((*d* + 1) / *d*). Newcomb noted this as a curiosity, and in the face of a general lack of interest, it was quickly forgotten.

In 1938, a physicist at General Electric name Frank Benford noticed the same pattern. Fascinated by this discovery, Benford set out to see exactly how well numbers from the real world correspond to the law. He collected 20,229 numbers from a variety of data sources including baseball statistics, areas of river catchments, and the addresses of the first 342 men listed in the book *American Men of Science*. The table at the top of the next page shows his observations. All these seemingly unrelated sets of numbers followed the same first-digit probability pattern as the worn pages of Newcomb's logarithm tables suggested. About 30% of the numbers began with a 1, 18% with a 2, and so on. His analysis was evidence for the existence of the law, but Benford was unable to explain quite why this should be so.

		First Digit Frequency								
Title	Sample Size	1	2	3	4	5	6	7	8	9
Area of Rivers	335	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1
Population	3259	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2
Constants	104	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6
Newspapers	100	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0
Specific Heat	1389	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1
Pressure	703	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7
H.P. Lost	690	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6
Mol. Wgt.	1800	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2
Drainage	159	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9
Atomic Wgt.	91	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5
Design	560	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6
Reader's Digest	308	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2
Cost Data	741	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1
X-Ray Volts	707	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8
Am. League	1458	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0
Blackbody	1165	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4
Addresses	342	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0
Death Rate	418	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1
Average	**1011**	**30.6**	**18.5**	**12.4**	**9.4**	**8.0**	**6.4**	**5.1**	**4.9**	**4.7**

Probability predictions are often surprising. In the case of the coin-tossing experiment, Dr. Hill wrote in *American Scientist* magazine, "a quite involved calculation" revealed a surprising probability. It showed, he said, that the overwhelming odds are that at some point in a series of 200 coin tosses, either heads or tails will come up six or more times in a row. Most fakers don't know this and avoid guessing long runs of heads or tails, which they mistakenly believe to be improbable. At just a glance, Dr. Hill can see whether or not a student's 200 coin-toss results contain a run of six heads or tails; if they don't, the student is branded a fake.

One of the experts putting Benford's Law to practical use is Dr. Mark Nigrini, an accounting consultant affiliated with the University of Kansas and Southern Methodist University in Dallas. Dr. Nigrini gained recognition a few years ago by applying a system he devised based on Benford's Law to some fraud cases in Brooklyn. The idea underlying his system is that if the numbers in a set of data like a tax return more or less match the frequencies and ratios predicted by Benford's Law, the data are probably honest. But if a graph of such numbers is markedly different from the one predicted by Benford's Law, he said, "I think I'd call someone in for a detailed audit."



(From "The First-Digit Phenomenon" by T. P. Hill, *American Scientist*, July-August 1998)

Benford's law can be used to test for fraudulent or random-guess data in income tax returns and other financial reports. Here the first significant digits of true tax data taken by Mark Nigrini from the lines of 169,662 IRS model files follow Benford's law closely. Fraudulent data taken from a 1995 King's County, New York, District Attorney's Office study of cash disbursement and payroll in business do not follow Benford's law. Likewise, data taken from the author's study of 743 freshmen's responses to a request to write down a six-digit number at random do not follow the law. Although these are very specific examples, in general, fraudulent or concocted data appear to have far fewer numbers starting with 1 and many more starting with 6 than do true data.

Some of the tests based on Benford's Law are too complex for hand calculations, but others are surprisingly simple. Just finding too few ones and too many sixes in a sequence of data to be consistent with Benford's Law is sometimes enough to arouse suspicion of fraud.

One of the earliest experiments Dr. Nigrini conducted with his Benford's Law program was an analysis of President Clinton's tax return. Dr. Nigrini found that it probably contained some rounded-off estimates rather than precise numbers, but he concluded that his test did not reveal any fraud.

The fit of number sets with Benford's Law is not infallible.

"You can't use it to improve your chances in a lottery," Dr. Nigrini said. "In a lottery someone simply pulls a series of balls out of a jar, or something like that. The balls are not really numbers; they are labeled with numbers, but they could just as easily be labeled with the names of animals. The numbers they represent are uniformly distributed, every number has an equal chance, and Benford's Law does not apply to uniform distributions."

Another problem Dr. Nigrini acknowledges is that some of his tests may turn up too many false positives. Various anomalies having nothing to do with fraud can appear for innocent reasons.

For example, the double digit 24 often turns up in analyses of corporate accounting, biasing the data, causing it to diverge from Benford's Law patterns and sometimes arousing suspicion wrongly, Dr. Nigrini said. "But the cause is not real fraud, just a little shaving. People who travel on business often have to submit receipts for any meal costing $25 or more, so they put in lots of claims for $24.90, just under the limit. That's why we see so many 24's."

Dr. Nigrini said he believes that conformity with Benford's Law make it possible to validate procedures developed to fix the Year 2000 problem -- the expectation that many computer systems will go awry because of their inability to distinguish the year 2000 from the year 1900. A variant of his Benford's Law software already in use, he said, could spot any significant change in a company's accounting figures between 1999 and 2000, thereby detecting a computer problem that might otherwise go unnoticed.

"I foresee lots of uses for this stuff, but for me its just fascinating in itself," Dr. Nigrini said. "For me, Benford is a great hero. His law is not magic, but sometimes it seems like it."

Dow Illustrates Benford's Law

To illustrate Benford's Law, Dr. Mark J. Nigrini offered this example:

"If we think of the Dow Jones stock average as 1,000, our first digit would be 1.

"To get to a Dow Jones average with a first digit of 2, the average must increase to 2,000, and getting from 1,000 to 2,000 is a 100 percent increase. "Let's say that the Dow goes up at a rate of about 20 percent a year. That means that it would take five years to get from 1 to 2 as a first digit.

"But suppose we start with a first digit 5. It only requires a 20 percent increase to get from 5,000 to 6,000, and that is achieved in one year.

"When the Dow reaches 9,000, it takes only an 11 percent increase and just seven months to reach the 10,000 mark, which starts with the number 1. At that point you start over with the first digit a 1, once again. Once again, you must double the number -- 10,000 -- to 20,000 before reaching 2 as the first digit.

"As you can see, the number 1 predominates at every step of the progression, as it does in logarithmic sequences."

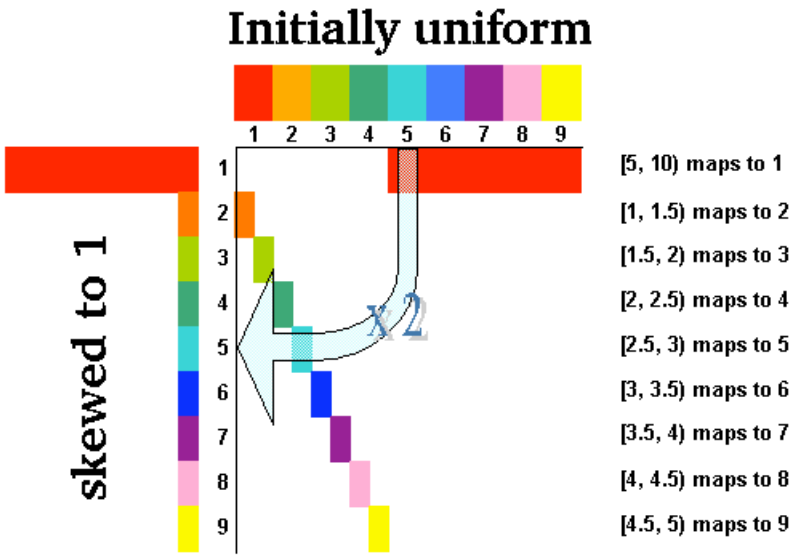
Deriving Benford's Law

The first step towards explaining this curious relationship was taken in 1961 by Roger Pinkham, a mathematician from New Jersey. Pinkham's argument was this:

Suppose that there really is a law of digit frequencies. If so, then that law should be universal; whether you measure prices in dollars, dinar, or drakma, whether you measure lengths in cubits, inches, or meters, the proportions of digit frequencies should be the same. In other words, Pinkham was saying that the distribution of digit frequencies should be *scale invariant*. Stated another way -- there's no reason why only one measurement scale, the one we happen to choose, should be the "correct one." The distribution of first significant digits should not change when every number is multiplied by a constant factor.

Many people have the intuition that each of the digits 1-9 are equally likely to appear as the leading digit of any number. Let's suppose this is the case and see what happens with a set of accounts that are to be converted from sterling to the euro at the fictional rate of 2 euros to the pound.

It's fairly easy to work out what will happen by looking at each digit in turn. If the first significant digit is 1, then multiplying by 2 will yield a new first digit of 2 or 3 with equal probability. But if the leading digit is 5, 6, 7, 8, or 9, the new first digit must be 1. It turns out that in the new set of accounts, a first digit of 1 is 10 times more likely than any other leading digit!



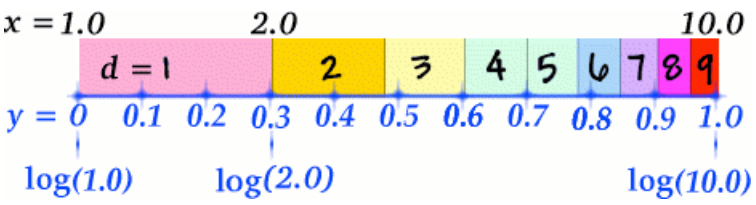
Our intuition has failed us – the original uniform distribution is now heavily skewed towards the digit 1. So if scale invariance is correct, the uniform distribution is the wrong answer.

So what does scale invariance of the distribution of the leading digit really mean? It means that if we multiply all our numbers by an arbitrary constant (as we do when we change from dollars to yen, or feet to meters), then the distribution of the leading digit frequencies should remain unchanged.

Since we are interested in the distribution of leading digits, it makes sense to express numbers in scientific notation: $x \times 10^n$. This is possible for all numbers except zero. The leading digit, *d*, is then simply the first digit of *x*. We can easily derive a scale invariant distribution for *d* once we have found a scale invariant distribution for *x*.

If a distribution for *x* is scale invariant, then the distribution of $y = \log x$ should remain unchanged when we *add* a constant value to *y*. This is true because we would be *multiplying* *x* by some constant *a* and $\log ax = \log a + \log x = \log a = y$.

Now the only probability distribution on *y* in [0,1) that will remain unchanged after the addition of an arbitrary constant to *y* is the uniform distribution. To convince yourself of this, think about the shape of the probability density function for the uniform distribution.



In the previous figure, *y* is uniformly distributed between $\log(1) = 0$ and $\log(10) = 1$.

If we want to find the probability that *d* is 1, we have to evaluate: $P(d = 1) = P(1 \le x \le 2) = P(0 \le y \le \log 2)$

To find this, we calculate the integral: $\int_0^{\log 2} 1 dy = \log 2 \approx .301$.