The Illusion of Randomness: Detecting Fraud using Benford's Law

Imagine that you have a fair coin, and you flip it. It lands on heads. Now what do you think would happen if you flip it again?

Some of you might be thinking that the first flip has no impact on the second (in other words, the two events are independent of each other), and you would be right. However, many people would predict that the coin would land on tails, almost as if tails is somehow 'due' to come next. This is a classic example of the Gambler's Fallacy.

The Gambler's Fallacy is a classic example of the fact that our intuition about randomness is often unreliable. In many ways, this isn't surprising: humans are naturally inclined to recognise patterns, as it is very useful in everyday life. However, this tendency comes at a cost: even when people try to produce 'randomness', they often introduce subtle patterns without realising it. This can have important consequences: in situations where fraudsters attempt to fabricate seemingly random data, they may believe that they're generating convincing values, but they end up with figures containing hidden structures that differ from genuine data. Recognising these hidden patterns can provide a vital clue to identifying that the data has been manipulated.

Imagine that you're investigating financial fraud, and you've been given a large dataset where some of the values might have been artificially generated. At first glance, the numbers may seem entirely ordinary: there are no obvious repetitions, missing entries, or visible signs of manipulation that suggest anything unusual. This is where the problem becomes difficult to detect. Fraudulent data is often designed to look random, so a simple visual inspection is unlikely to reveal anything abnormal. In fact, fraudsters may even use random number generators to avoid accidentally introducing noticeable patterns. However, this is where mathematics comes to the rescue. In several real-world datasets, a consistent pattern emerges within the distribution of leading digits in numbers. Rather than being evenly distributed, smaller digits appear as leading digits far more often than larger ones. This phenomenon is known as Benford's Law.

Let's look at Benford's Law (also known as the First Digit Law) from a mathematical perspective. Any number 'n' can be written in standard form as $a \times 10^b$, where $1 \leq a < 10$, and b is an integer.

$$n = a \times 10^{b}$$

Taking the base-10 logarithm and rearranging gives:

$$log_{10}(n) = b + log_{10}(a)$$

The integer part of $\log_{10}(n)$, b, doesn't affect the leading digit of n, but the fractional part, $\log_{10}(a)$, does. Therefore the distribution of leading digits depends on how this fractional part is distributed.

If we assume that these fractional parts are uniformly distributed between 0 and 1, which is often approximately true for many real-world datasets, then the probability that a number has leading digit 'd', where $d \in \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$, depends on the size of the interval that produces that digit. In particular, a number begins with the digit 'd' when:

$$log_{10}(d) \leq log_{10}(a) < log_{10}(d + 1)$$

Therefore, the probability of the leading digit being 'd' is:

$$P(d) = log_{10}(d + 1) - log_{10}(d) = log_{10}(\frac{d+1}{d}) = log_{10}(1 + \frac{1}{d})$$

This is exactly what Benford's Law states.

This result shows why lower numbers are more likely to occur as leading digits (for example, in many datasets, the chance of the leading digit being 1 is about 30.1%, but the probability of the leading digit being 9 is only about 4.6%)[1].

d	P(d) ~ 1 d.p.
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%

At first glance, this result may seem surprising. It is natural to assume that each digit from 1 to 9 would be equally likely to appear as the first digit of a number. In other words, we might expect each digit to occur as the leading digit about one ninth of the time (around 11.1%). However, this result is more intuitive when we consider that many real-world processes don't grow linearly. Instead, they tend to grow multiplicatively, with values increasing by proportions rather than fixed steps, such as with compound interest. This type of exponential growth is what leads to the distribution of leading digits described by Benford's Law.

This law appears in many purely mathematical settings, such as in exponential sequences like 2^x, and the Fibonacci numbers, which suggests that the law is deeply rooted in the structure of number systems themselves.

Because Benford's Law is widely applicable across many real-world contexts, but not widely recognised, it can be a very useful benchmark for detecting potential fraud (at least until it becomes a more widely known law). Data that has been manipulated is unlikely to follow the same trend as what Benford's Law states, and may instead have a more even spread of leading digits. As a result, significant deviations from the expected pattern stated by the model may suggest that some of the values in the dataset have been artificially constructed[2].

But this raises an important question: when can Benford's Law be used?

Its reliability strongly depends on the nature of the dataset. In general, the datasets that Benford's Law can best be applied to are large datasets that span across multiple orders of magnitude and contain values that aren't intentionally chosen to fit a pattern, values that arise naturally from real-world activity. In the case of many financial datasets, all of these conditions are naturally satisfied, making it a reasonably reliable tool to help in detecting anomalies that could indicate fraudulent activity.

However, there are also important limitations to the use of Benford's Law. For example, it can't be used with datasets where values are manually assigned, such as with identification codes or invoice numbers, because those values don't arise from natural processes, and therefore have no reason to follow Benford's Law. Furthermore, if the datasets only involve a narrow range of values (e.g. every number in the dataset is between £10 and £20), Benford's Law is unlikely to emerge, even if the values in them are completely legitimate.

There is also one crucial problem: in theory, it is possible to deliberately construct data that mimics the expected leading digit distribution of Benford's Law. A fraudster who's aware of Benford's Law could generate or adjust values to create a dataset that appears to follow the expected pattern. As a result, Benford's Law can only be used to verify if the data behaves as expected, rather than confirming whether the data is truthful or not. Its reliability depends on if we can assume that the data hasn't been manipulated in order to follow the expected leading digit distribution.

So, in practice, is it simple to fabricate data to avoid detection? In reality, it is definitely not straightforward. Datasets tend to be very large, making it difficult to consistently manipulate values so that they follow Benford's Law. Moreover, in forensic accounting, Benford's Law isn't used in isolation, but alongside a variety of other statistical tools, improving its overall detection capability.

For example, Zipf's Law states that in naturally occurring datasets, the most frequent value occurs about twice as often as the second most frequent value, about thrice as often as the third most frequent value and so on: it could indicate money laundering if it detects unusual patterns in wire transfers (such as a high frequency of unique transactions). Methods such as Z-Scores and Chi-Square Tests can measure how far a dataset differs from its expected behaviour. Together, these tools can all be used to build a more complete analysis of whether a dataset is genuine or not.

In conclusion, human intuition quite often fails to replicate what true randomness looks like. However, this can be useful when it comes to fraud detection. Benford's Law is very effective in identifying unusual data by examining the distribution of leading digits to reveal a pattern. However, there are limitations to this approach, as it can only detect when data appears suspicious, rather than proving that the data is fraudulent. This highlights a broader idea within mathematics: it doesn't tell us what's true – it tells us what's worth questioning.

Bibliography:

1.https://www.sciencedirect.com/science/article/pii/S2667325823000043
2.https://www.jstor.org/stable/27643897?searchText=benford%27s+law+financial+fraud&searchUri=%2Faction%2FdoBasicSearch%3FQuery%3Dbenford%2527s%2Blaw%2Bfinancial%2Bfraud%26so%3Drel&ab_segments=0%2Fbasic_search_gsv2%2Fcontrol&refreqid=fastly-default%3A84a2300baaa597cb342522c3b9c2d3e5&seq=1