Benford's Law

Owen BARRON[*]

[*] Coláiste an Spioraid Naoimh, University College Cork Maths Enrichment Program

I encountered Benford's Law for the first time at a young age while reading through a book on mathematical curiosities. The simplicity and generality of the law made it fascinating to me – especially since it appeared completely impossible at first sight.

Benford's Law is an interesting and at first counter-intuitive statistical phenomenon that occurs in almost all natural data sets that span multiple orders of magnitude. It states that the first digit of items in these sets is much more likely to be small (e.g. 1 or 2) than big (e.g. 8 or 9). The expected distribution is staggeringly weighted towards these small digits. According to the law, 1 occurs as the first digit of about 30% of entries in data sets, whereas 9 appears first in only 5%!

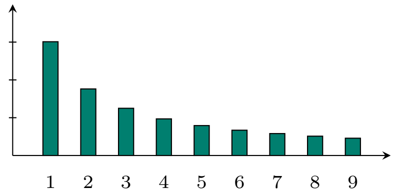

Fig. 1. Frequencies defined by the Benford's Law

Precisely, Benford's Law states that the frequency with which a digit d occurs in a base-10 data set spanning some orders of magnitude of powers of 10 is roughly equal to $\log_{10}(d+1) - \log_{10}(d)$ (Fig. 1). The spanning some orders of magnitude condition is deceptively important – data sets not following this condition rarely follow the law. As an example, the number of pages in books do not for the clear reason that most page counts are somewhere between 200 and 600.

It's difficult to overestimate how broadly this law applies: some data sets that follow it in the real world are the heights of the hundred tallest buildings or the lengths of rivers. Examples even extend to the purely mathematical, such as the Fibonacci series, the series of powers of 2, and even the series obtained from alternating between multiplication by 2 and 3! After hearing about the law for the first time, I was slightly bemused – how could something so blatantly asymmetric hold true in such a generalised sense? This article will aim to provide an intuitive if not formal explanation for why the law holds, as well as some history behind it and surprising places it can be applied in real life.

The law was first discovered in 1881 by Simon Newcomb, a Canadian-American astronomer. He observed that the earlier pages of books of logarithms, used for calculations were much more worn out than those later on. He conjectured that this was due to the data sets scientists were performing calculations on tending towards having numbers with lower starting digits. Newcomb published a brief note on the phenomenon including the theoretical values of probabilities and a short, informal argument explaining why it was true in the American Journal of Mathematics, but it gained minimal traction.

More than fifty years later, in 1938, the law was rediscovered independently by its namesake – Frank Benford. Benford was working in research physics when he discovered the pattern in the same manner that Newcomb had. However, he took his investigation of the law a level deeper and gathered sets of data with 20,000 total items as examples. He used these as evidence in the paper he published in the Proceedings of the American Philosophical Society, aptly titled 'The Law of Anomalous Numbers'. Unfortunately this name did not stick, and the law was instead called after Benford, giving us another example of Stigler's Law in the mathematical world. world

Stigler's Law states that discoveries are relatively rarely named after the people who made them. One example of this is Stigler's Law itself.

The reader may be forgiven for some skepticism about all of the above at this point – but what may at first seem like an impossible phenomenon luckily has a simple and intuitive explanation. Say we have a set of numbers spanning multiple orders of magnitude. We would expect the data points to be roughly evenly distributed among these orders – e.g there are roughly the same number of items in the interval $[100, 1000]$ as $[1000, 10\,000]$. On the logarithmic scale those two intervals have the same length $\log_{10} 10 = 1$. By extrapolation we may expect that for any two segments of equal length on the logarithmic scale the number of items falling in either of them is roughly the same. In other words, we expect the data points to be evenly spread on the logarithmic scale. The set of

We refer the reader to the article *Od mnożenia do dodawania* (Δ^{01}_{25}) for further interesting information on the logarithmic scale

numbers whose base 10 representation starts with a digit d is a disjoint union of sequences of length $\log_{10}(d+1) - \log_{10}(d)$, which leads us to the probability given by the Benford's law.

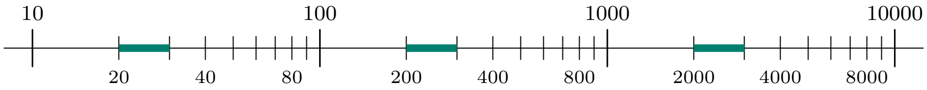

Fig. 2. Coloured segments of length $\log_{10}(3) - \log_{10}(2)$ correspond to having '2' as the first digit

Let us note that the above reasoning does not constitute a full *proof of correctness* of Benford's Law. Theoreticians may argue that a „uniform distribution" does not exist over the entire infinite logarithmic scale, while practitioners may ask why we actually are focusing on a uniform distribution in the logarithmic scale — whether this is a law of nature or if it also follows from mathematical theorems of the kind found in the central limit theorem. For readers who find this explanation unsatisfactory, we recommend Ted Hill's article *A Statistical Derivation of the Significant-Digit Law*, published in *Statistical Science* in 1995.

Unlike many high-level mathematical theorems, there are a surprising number of situations in the real world where direct application of the law becomes extremely useful. The most important of these applications occur in the field of fraud detection, where it is often used as a preliminary test to find faults. Accounting books normally follow Benford's Law, as would be expected from a roughly random distribution over several orders of magnitude. However, when random numbers are generated by computer or by hand in fabricated accounts, one would expect them to instead have evenly distributed digits. After first hearing about Benford's Law, financial investigator Darrell D. Dorrell immediately began applying it to cases he was working on. This led to the successful conviction of Wesley Rhodes, a financial advisor who embezzled millions of dollars in funds from his investors. The law is regularly used as a first indicator or red flag of financial fraud – if the figures in checkbooks do not follow the law, it is likely some kind of anomaly is at play.

One other very interesting use was the uncovering of a hidden bot network on Twitter. Jennifer Golbeck noticed in 2015 that for the majority of users, the number of followers that their followers have adheres to Benford's Law. However, a small percentage of accounts investigated did not adhere to the pattern. These 170 accounts were flagged and investigated further, by examining their followers and post history. Out of all 170 only 2 seemed to belong to legitimate users. The rest of the accounts all had followers among the other accounts, and clearly automated or otherwise suspicious posts.

However, the law has strong failings when it comes to some other attempted uses, especially in election fraud. There is a simple reason for this: electoral districts normally have similar populations. Thus if one candidate expects to receive a certain percentage of votes in each of these districts, their first digit distribution will be confined to a range not necessarily following Benford's Law. Similarly for financial fraud, if a company sells a large number of a product that has a specific price, the digit distribution in accounts will be weighted towards the first digit of this product. In the 2020 US presidential race, conspiracists noticed that Joe Biden's vote counts in some areas did not follow the law. They were quick to raise this issue as an indicator of a rigged election, but for the reasons above they were proven to be mistaken in doing so.

Somewhat surprisingly, a much weaker version of the law holds for the second digit of numbers in data sets as well – in these cases, the difference in occurrences between 0 and 9 as the second digit is only about 3% – but this is still recognizable in large enough sets. There is in fact a distribution in the general case for the n^{th} digit, although it becomes flatter and flatter as n increases. In the application of Benford's Law to elections, it is this generalisation in combination with the first digit law that allows a decision on whether or not fraud exists to be much more reliable. The second digit of a number will clearly be less affected in general by the similarity of electoral district populations than the first digit.

The n^{th} digit is not the only generalisation of the law. It also holds when the data points are converted into any other base, and even when the data points are converted from one unit into another. In my research on the topic, I discovered other similarly surprising results too – if you are interested, Zipf's Law on the frequency of words in texts is fascinating. I am still surprised every time the law shows up in my day-to-day life – if you start searching for it, you may begin to see examples everywhere you go!