146. Benford's law – and astrometry

A CURIOUS MATHEMATICAL property of many collections of numbers, such as mathematical tables and naturally occurring data, is that the leading digits are not uniformly distributed, but are significantly skewed toward smaller values. This was first noted by US astronomer–mathematician Simon Newcomb (1881). Rediscovered by Frank Benford (1938), it is generally known as Benford's Law.

Newcomb's paper, in the American Journal of Mathematics of 1881, opens with the statement: *'That the ten digits do not occur with equal frequency must be evident to anyone making much use of logarithmic table, and noticing how much faster the first pages wear out than the last ones. The first significant figure is oftener 1 than any other digit, and the frequency diminishes up to 9.'*

Newcomb was, incidentally, professor of mathematics in the United States Navy and at Johns Hopkins University, director of the Nautical Almanac Office, and made major contributions to the study of planetary and lunar motion, the recalculation of the major astronomical constants, the measurement of the speed of light, and the interpretation of the Earth's Chandler wobble.

AS FORMULATED BY NEWCOMB, and in the particular case of the first significant decimal (base 10) digits, the 'law' corresponds to a distribution expected if the logarithm of the numbers are uniformly and randomly distributed. Then the leading digit d (for $d = 1, 2, \ldots 9$) occurs with probability $P(d) = \log(d+1) - \log(d) = \log(1 + d^{-1})$. Then $d = 1$ occurs in 30.1% of the cases, $d = 2$ in 17.6%, and so on up to just 4.6% for $d = 9$.

The law is most accurate when values are distributed across many orders of magnitude. It has parallels with Zipf's law, used in code breaking and quantitative linguistics, in which the word frequency in a text or language corpus is inversely proportional to the word rank.

Benford tested the law on various disparate data sets, including the surface areas of 335 rivers, the sizes of 3259 US community populations, 104 physical constants, and 1800 molecular weights. Others have examined its applicability to widely ranging data since.

THERE IS a substantial mathematical literature that continues to examine its wider applicability and significance (e.g. Berger & Hill, 2020; Burgos & Santos, 2021), including the substantial Benford Online Bibliography of articles and other related resources. Like some other broad principles applied to natural data (e.g. that many data sets are well approximated by a normal distribution) there are explanations that cover many cases where Benford's law applies, though there are others that seem to challenge a straightforward understanding.

In physics, it has been applied to complex atomic spectra (Pain, 2008), full width of hadrons (Shao & Ma, 2009), and half-life times for both α decay (Buck et al., 1993), and β decay (Dong-Dong et al., 2009).

As an example from number theory, among the first billion powers of 2, exactly 301 029 995 begin with the digit 1, while the Benford law prediction for this count is $10^9 \log_{10} 2 = 301\,029\,995.66$ (Cai et al. 2020).

It has also been exploited in practical applications, including identifying suspicious accounting and tax returns, voter fraud, and altered digital images, where fabricated data deviates from the underlying expectations.

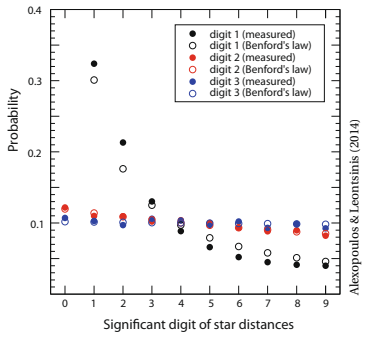
IN ASTRONOMY, it was first studied in the context of the distances of galaxies and stars by Alexopoulos & Leontsinis (2014). They concluded that *'the distances of galaxies follow the first digit law reasonable well, and that the star distances agree very well with the first, second and third significant digit'*.

Hill & Fox (2016) gave a theoretical explanation, based on Hubble's law and the mathematical properties of Benford's law, as to why galaxy distances might be expected to follow Benford's law, and even argued that *'the recent empirical observations may be viewed as independent evidence of the validity of Hubble's law'*.

Shukla et al. (2017) suggest that some (but not all) Kepler planet properties, such as mass, semi-major axis, and period, also follow this distribution. This was confirmed by Melita & Miraglia (2021), although radii fail. Mamidipaka & Desai (2023) found that the dispersion measures of pulsars and Fast Radio Bursts also fail.

Is Benford's law of any relevance to Gaia? I will admit that, applied to these topics in astronomy, I waver between whether it is a statement of the (relatively) obvious, or truly encapsulates something more profound. But meanwhile I will simply report on what has been published in the literature!

As I mentioned previously, the first attempt to compare star distances with the predictions of Benford's law was made by Alexopoulos & Leontsinis (2014). They used the somewhat *ad hoc* HYG database (Hipparcos–Yale–Gliese) compiled by amateur astronomer David Nash. As shown here, they found that, for their 115 256 star *distances* (computed simply as the reciprocal of the published parallax), the first, and especially the second and third significant digits follow the probabilities predicted by Benford's law rather well.

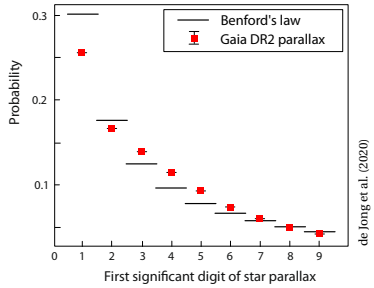


With the availability of Gaia DR2, de Jong et al. (2020) showed that the 1.3 billion astrometric *parallaxes* follow Benford's law even closer. Stars with a parallax starting with digit 1 are five times more numerous than stars with a parallax starting with the digit 9.

But they reached a different conclusion for the stellar *distances* (determined as the reciprocal of the parallax), and I need to give some background to explain why this might be so, and why it might be of interest.

When the relative parallax error exceeds 10–20%, the reciprocal of the observed parallax is a sub-optimal estimate of the distance. This is evident from the fact that Gaia (and Hipparcos) measures the *parallax* with some associated standard error. But this symmetrical error in parallax transforms into an asymmetric probability distribution in distance. Improved *distance* estimates, exploiting the use of prior assumptions, have been variously proposed (e.g. Bailer-Jones, 2015). In the case of Gaia DR2, improved distance estimates have been based on prior information about the Galaxy (Bailer-Jones et al., 2018; Luri et al., 2018), or by including additional photometric data (Anders et al., 2019).

Even taking into account these improved estimates, as well as considering possible offsets in the parallax zero-point, they showed that the *distances* estimated from Gaia do not follow Benford's law, although distances with small starting digits are still more abundant.



Moreover, using realistic simulations of the stellar content of the Galaxy (Robin et al., 2012), they argued that the distances would *not* be expected to follow Benford's law, essentially because the combination of the luminosity function of the Milky Way, and Gaia's selection function, results in a bi-modal distance distribution, corresponding to nearby dwarfs in the Galactic disk, and more distant giants in the Galactic bulge.

But the fact that the observed Gaia DR2 parallaxes follow Benford's law, at least tolerably, while the true distances derived from the parallaxes do not, is, they conclude, the most intriguing result of their study.

I will add an aside on the subject of 'misnamed theorems'. Benford's law, having been discovered by Simon Newcomb, is therefore an example of Stigler's 'law of eponymy', which asserts (acerbically) that no scientific discovery is named after its original discoverer!

The wiki entry for Stigler's law gives many examples. Amongst those in astronomy are: Bode's law first stated by Johann Titius; the Cassegrain reflector known to both Cavalieri and Mersenne; Dyson spheres which Dyson himself credited to Olaf Stapledon; the Fermi paradox previously stated by Tsiolkovsky; the Gaussian distribution introduced by de Moivre; Halley's comet observed since at least 240 BCE; Hubble's law derived by Georges Lemaître; Kapteyn's star previously catalogued by B.A. Gould; the Oort cloud first postulated by Ernst Öpik; and Olbers' paradox formulated by Kepler.

Amusingly, Stephen Stigler himself, in an article in 1980, named sociologist Robert K. Merton as the discoverer of 'Stigler's law', thereby showing that it follows its own misnaming.

I will finish with two quotes from Simon Newcomb, which reveal the context in which astronomy was viewed by one of its foremost exponents more than a century ago. In 1888 he wrote: *'We are probably nearing the limit of all we can know about astronomy'.*

But his views changed towards the end of his life (Newcomb, 1903): *'What lies before us is an illimitable field, the existence of which was scarcely suspected ten years ago, the exploration of which may well absorb the activities of our physical laboratories, and of the great mass of our astronomical observers and investigators for as many generations as were required to bring electrical science to its present state'.*