Benfords' Law: Gaia parallaxes and distances

Jos de Bruijne[1], Jurjen de Jong[1,2,3], Joris De Ridder[2]

[1]European Space Agency, Directorate of Science, Science Support Office (SCI-S), ESTEC, Noordwijk, The Netherlands
[2]Instituut voor Sterrenkunde, Leuven, Belgium
[3]Matrixian Group: The Data Science Company, Amsterdam, The Netherlands

Benford's Law states that the frequency distribution of significant digits of data sets representing natural phenomena covering a large dynamic range such as terrestrial river lengths and mountain heights is non-uniform, with a strong preference for small numbers. As an example, Benford's Law states that 1 appears as the leading significant digit 30.1% of the time while 9 occurs as first significant digit for only 4.6% of the data points. Alexopoulos & Leontsinis (2014) demonstrated that the ~100,000 Hipparcos parallaxes, converted to distances by naive inversion, follow Benford's Law. We present an investigation into the intriguing question whether also the 1.3 billion Gaia DR2 parallaxes, and the associated Bayesian-inferred distances, follow Benford's Law.

Benfords' Law

Benford's Law (1938) has the following probability distribution $P(d)$ for the first significant digit d:

$$P(d) = \log_{10}\left(1+\frac{1}{d}\right) \text{ for } d \in \{1,2,3,...,9\}$$

	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$	$d=7$	$d=8$	$d=9$
$P(d)$	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Benford's Law only applies to non-truncated datasets that span several orders of magnitude such as river lengths or mountain heights.

If datasets follow Benford's Law, then they must be scale- and base-invariant (e.g., neither conversion from parsec to lightyear nor conversion from base 10 to base 6 should matter). Therefore, if a dataset follows Benford's Law, then also the "inverse" dataset follows it.

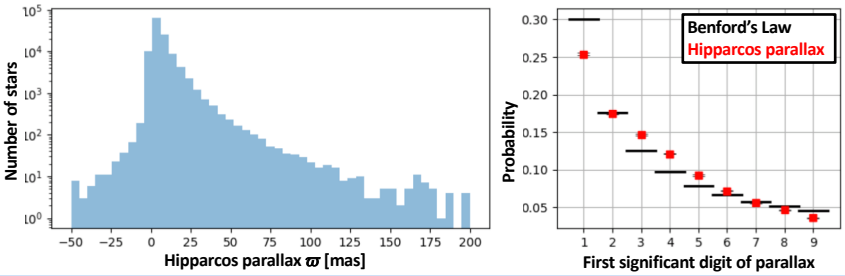
Leading digit	meters		feet		In Benford's law
	Count	%	Count	%	
1	26	43.3%	18	30.0%	30.1%
2	7	11.7%	8	13.3%	17.6%
3	9	15.0%	8	13.3%	12.5%
4	6	10.0%	6	10.0%	9.7%
5	4	6.7%	10	16.7%	7.9%
6	1	1.7%	5	8.3%	6.7%
7	2	3.3%	2	3.3%	5.8%
8	5	8.3%	1	1.7%	5.1%
9	0	0.0%	2	3.3%	4.6%

Example: Benford's Law and the 60 tallest structures in the world (from Wikipedia).

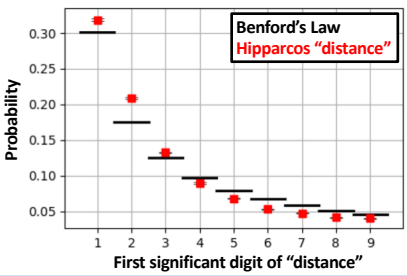
Hipparcos

Alexopoulos & Leontsinis (2014) investigated the first-significant digit distribution of Hipparcos "distances" and saw Benford's Law.

The Hipparcos parallaxes span several orders of magnitude (left) and their first-significant digit distribution resembles Benford's Law (right). Inverse parallaxes ("distances") should then also show Benford's Law ...



Simply following Alexopoulos & Leontsinis (2014), we derive "distances" by parallax inversion (when $\varpi > 0$ mas). Since small, positive parallaxes ($0 < \varpi < 1$ mas) are abundant, many stars are placed beyond 1 kpc. The "distances" span several orders of magnitude and resemble Benford's Law ...



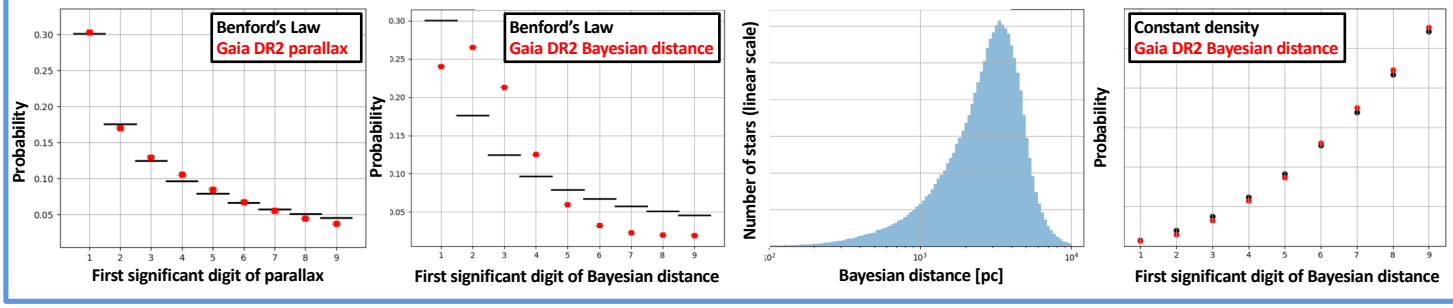
Gaia DR2

We investigate the first-significant digit distribution of the Gaia DR2 parallaxes (Gaia Collaboration et et al. 2016, 2018) and the Bayesian-inferred distances from Bailer-Jones et al. (2018).

The Gaia DR2 parallaxes nicely follow Benford's Law. It must be noticed, however, that applying the recommended quality filters (Lindegren et al. 2018) and/or filtering out negative parallaxes does influence the data at the level of a few hundredths in probability.

The Gaia DR2 Bayesian distance estimates from Bailer-Jones et al. (2018), however, do not follow Benford's Law (left). Most stars in Gaia DR2 are located at 2-3 kpc from the Sun (right). This, clearly, is a natural result of the exponential-disk-like Milky-Way prior adopted in the Bayesian framework. In short, the first-significant digit distribution of the Gaia DR2 distance estimates does not follow Benford's Law but is the result of the interplay of the Milky-Way luminosity function and extinction distribution, the Galactic-disk structure, and the Gaia DR2 faint (completeness) limit.

For stars within 100 pc – and actually this is true up to ~800 pc – the first-significant digit distribution of Gaia DR2 distance estimates is fully compatible with the expected distribution for stars that are distributed in a sphere with constant density.



References

Alexopoulos and Leontsinis, 2014, JApA, 35, 639 (https://doi.org/10.1007/s12036-014-9303-z)
Bailer-Jones et al., 2018, AJ, 156, 58 (https://doi.org/10.3847/1538-3881/aacb21)
Benford, 1938, Proc. Am. Phil. Soc., 78, 551 (https://www.jstor.org/stable/984802)
Gaia Collaboration et al., 2016, A&A, 595, A1 (https://doi.org/10.1051/0004-6361/201629272)
Gaia Collaboration et al., 2018, A&A, 616, A1 (https://doi.org/10.1051/0004-6361/201833051)
Lindegren et al., 2018, A&A, 616, A2 (https://doi.org/10.1051/0004-6361/201832727)

This work has made use of data from the European Space Agency (ESA) mission Gaia (https://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement.