A&A 642, A205 (2020)
https://doi.org/10.1051/0004-6361/201937256
© ESO 2020

Astronomy & Astrophysics

Benford's law in the *Gaia* universe

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

[1] Science Support Office, Directorate of Science, European Space Research and Technology Centre (ESA/ESTEC), Keplerlaan 1, 2201 AZ Noordwijk, The Netherlands
e-mail: jurjendejong93@gmail.com
[2] Institute of Astronomy, KU Leuven, Celestijnenlaan 200D, 3001 Leuven, Belgium
[3] Matrixian Group, Transformatorweg 104, 1014 AK Amsterdam, The Netherlands

Received 4 December 2019 / Accepted 18 August 2020

ABSTRACT

Context. Benford's law states that for scale- and base-invariant data sets covering a wide dynamic range, the distribution of the first significant digit is biased towards low values. This has been shown to be true for wildly different datasets, including financial, geographical, and atomic data. In astronomy, earlier work showed that Benford's law also holds for distances estimated as the inverse of parallaxes from the ESA HIPPARCOS mission.

Aims. We investigate whether Benford's law still holds for the 1.3 billion parallaxes contained in the second data release of *Gaia* (*Gaia* DR2). In contrast to previous work, we also include negative parallaxes. We examine whether distance estimates computed using a Bayesian approach instead of parallax inversion still follow Benford's law. Lastly, we investigate the use of Benford's law as a validation tool for the zero-point of the *Gaia* parallaxes.

Methods. We computed histograms of the observed most significant digit of the parallaxes and distances, and compared them with the predicted values from Benford's law, as well as with theoretically expected histograms. The latter were derived from a simulated *Gaia* catalogue based on the Besançon galaxy model.

Results. The observed parallaxes in *Gaia* DR2 indeed follow Benford's law. Distances computed with the Bayesian approach of Bailer-Jones et al. (2018, AJ, 156, 58) no longer follow Benford's law, although low-value ciphers are still favoured for the most significant digit. The prior that is used has a significant effect on the digit distribution. Using the simulated *Gaia* universe model snapshot, we demonstrate that the true distances underlying the *Gaia* catalogue are not expected to follow Benford's law, essentially because the interplay between the luminosity function of the Milky Way and the mission selection function results in a bi-modal distance distribution, corresponding to nearby dwarfs in the Galactic disc and distant giants in the Galactic bulge. In conclusion, *Gaia* DR2 parallaxes only follow Benford's Law as a result of observational errors. Finally, we show that a zero-point offset of the parallaxes derived by optimising the fit between the observed most-significant digit frequencies and Benford's law leads to a value that is inconsistent with the value that is derived from quasars. The underlying reason is that such a fit primarily corrects for the difference in the number of positive and negative parallaxes, and can thus not be used to obtain a reliable zero-point.

Key words. astronomical databases: miscellaneous – astrometry – stars: distances

1. Introduction

Benford's law, sometimes referred to as the law of anomalous numbers or the significant-digit law, was put forward by Simon Newcomb in 1881 and later made famous by Frank Benford (Newcomb 1881; Benford 1938). The law states that the frequency distribution of the first significant digit of data sets representing (natural) phenomena covering a wide dynamic range such as terrestrial river lengths and mountain heights is non-uniform, with a strong preference for low numbers. As an example, Benford's law states that digit 1 appears as the leading significant digit 30.1% of the time, while digit 9 occurs as first significant digit for only 4.6% of data points taken from data sets that adhere to Benford's law. Although Benford's law has been known for more than a century and has received significant attention in a wide range of fields covering natural and (socio-)economic sciences (e.g. Berger & Hill 2015), a statistical derivation was only published fairly recently, showing that Benford's law is the consequence of a central-limit-theorem-like theorem for significant digits (Hill 1995b).

Alexopoulos & Leontsinis (2014) investigated the presence of Benford's law in the Universe and demonstrated that the ~118 000 stellar parallaxes from the ESA HIPPARCOS astrometry satellite (ESA 1997), converted into distances by inversion, follow Benford's law. In this paper, we extend this work and present an investigation into the intriguing question whether the ~1.3 billion parallaxes and the associated Bayesian-inferred distances that are contained in the second data release of the HIPPARCOS successor mission, *Gaia* (Gaia Collaboration et al. 2016, 2018), follow Benford's law as well. Moreover, we investigate the prospects of using Benford's law as tool for validating the *Gaia* parallaxes. The idea of using Benford's law as a tool for anomaly detection is not new: Nigrini (1996) described that Benford's law was used to detect fraud in income-tax declarations. We adapt this idea and investigate the effect of the parallax zero-point offset that is known to be present in the *Gaia* DR2 parallax data set (Lindegren et al. 2018) with the aim to determine whether Benford's law can be used to derive the value of the offset.

This paper is organised as follows. Section 2 presents a short overview of Benford's law. Section 3 summarises and discusses the HIPPARCOS-based study by Alexopoulos & Leontsinis (2014) that inspired this work. Section 4 presents our work, which is based on *Gaia* DR2. The effect of the parallax

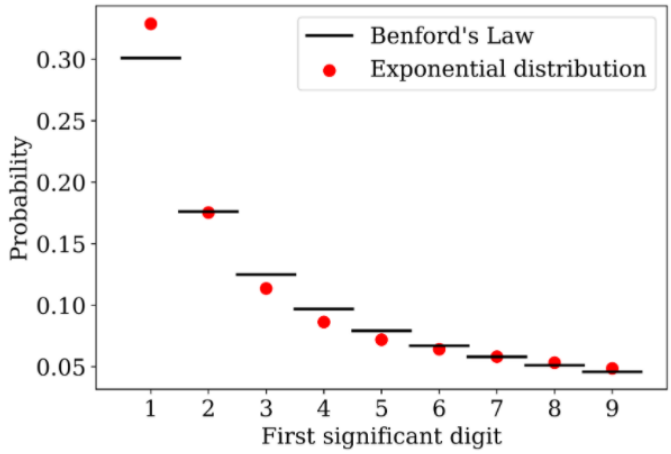

Fig. 1. Comparison of the frequency of occurrence of all possible values of the first significant digit ($d = 1, \ldots, 9$) between one million randomly drawn numbers from an exponential distribution (e^{-X}; red circles) and Benford's law (black, horizontal bars).

zero-point is discussed in Sect. 5, and a discussion and conclusions can be found in Sect. 6.

2. Benford's law

Benford's law is an empirical, mathematical law that gives the probabilities of occurrence of the first, second, third, and higher significant digits of numbers in a data set. In this paper, we limit ourselves to the first significant digit. We also investigated the second and third significant digits, but this did not yield additional insights into this study.

Every number $X \in \mathbb{R}_{>0}$ can be written in scientific notation as $X = x \cdot 10^m$, where $1 \leq x < 10$ and $x \in \mathbb{R}_{>0}$, $m \in \mathbb{Z}$. The quantity x is called the significand. The first significant digit is therefore also the first digit of the significand. This formulation allows us to define the first-significant digit operator D_1 on number X with a floor function:

$$D_1 X = \lfloor x \rfloor. \tag{1}$$

According to Benford's law, the probability for a first significant digit $d = 1, \ldots, 9$ to occur is

$$P(D_1 X = d) = \log_{10}\left(1 + \frac{1}{d}\right). \tag{2}$$

Benford's law states that the probability of occurrence of 1 as first significant digit ($d = 1$) equals $P(D_1 X = 1) = 0.301$. This probability decreases monotonically with higher numbers d, with $P(D_1 X = 2) = 0.176$, $P(D_1 X = 3) = 0.125$, down to $P(D_1 X = 9) = 0.046$ for $d = 9$ as first significant digit.

Figure 1 shows the first significant digit of randomly drawn numbers from an exponential distribution (e^{-X}) versus Benford's law. This shows that the exponential distributed data approaches Benford's law.

Benford's law is an empirical law. This means that there is no solid proof to show that a data set agrees with Benford's law. Nonetheless, the following conditions make it very likely that a data set follows Benford's law:

1. The data shall be non-truncated and rather uniformly distributed over several orders of magnitude. This can be understood through Eq. (2), which shows that on a logarithmic scale, the probability $P(D_1 X = d)$ is proportional to the space between d and $d + 1$. In other words: Benford's law results naturally if the mantissae (the fractional part) of the logarithms of the numbers are uniformly distributed. For example, the mantissa of $\log_{10}(2 \cdot 10^m) \approx m + 0.30103$ for $m \in \mathbb{Z}$, equals 0.30103. This is graphically illustrated in Fig. 2, in which we intuitively show that a distribution close to a uniform logarithmic distribution should obey Benford's law. In particular, the red bins with $d = 1$ occupy ~30% of the axis, compared to the ~5% length of the blue bins that contain numbers where $d = 8$. Even though the distribution is not perfectly uniform (the heights of the bins vary), the cumulative areas of all red and all blue bins are determined more by the (fixed) widths of the bins than by their heights, such that Benford's law is approximated when adding (i.e., averaging over) several orders of magnitude.

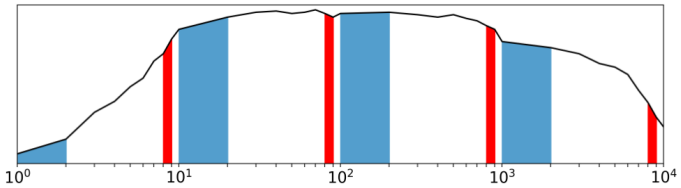

Fig. 2. Schematic example of a probability distribution of a variable that covers several orders of magnitude and that is fairly uniformly distributed on a logarithmic scale. The sum of the area of the blue bins is the relative probability that the first significant digit equals 1 ($d = 1$), while the sum of the area of the red bins is the relative probability that the first significant digit equals 8 ($d = 8$). Because the distribution is fairly uniform, i.e. the bin heights are roughly the same, the cumulative red and blue areas are foremost proportional to the fixed widths of the red and blue bins, respectively, such that numbers randomly drawn from this distribution will approximate Benford's law.

2. From the previous point, it follows intuitively that if a data set follows Benford's law, it must be scale invariant (see Appendix F.1). In particular, a change of units, for instance from parsec to light year when stellar distances are considered, should not (significantly) change the probabilities of occurrence of the first significant digit. This can be understood by looking at Fig. 2 and by considering a uniform logarithmic distribution for which the logarithmic property $\log Cx = \log C + \log x$ holds for variable $x \in \mathbb{R}_{>0}$ and constant (scaling factor) $C \in \mathbb{R}_{>0}$.

3. Hill (1995b) demonstrated that scale-invariance implies base-invariance (but not conversely). It therefore follows that if a data set follows Benford's law, it must be base invariant (see Appendix F.2). In particular, a change of base, for instance from base 10 as used in Eq. (2) to base 6, should not (significantly) change the probabilities of occurrence of the first significant digit in comparison to Benford's law.

For reasons explained in Appendix D, we used a simple Euclidean distance to quantify how well the distribution of the first significant digit of *Gaia* data is described by Benford's law. Although this sample-size-independent metric is not a formal test statistic, with associated statistical power, this limitation is acceptable in this work because we only use the Euclidean distance as a relative measure (see Appendix D for an extensive discussion).

3. Hipparcos

Alexopoulos & Leontsinis (2014) presented an assessment of Benford's law in relation to stellar distances. Unfortunately, the authors provide very limited information and just state that they used the distances from the HYG [Hipparcos-Yale-Gliese] database, which includes 115 256 stars with distances reaching

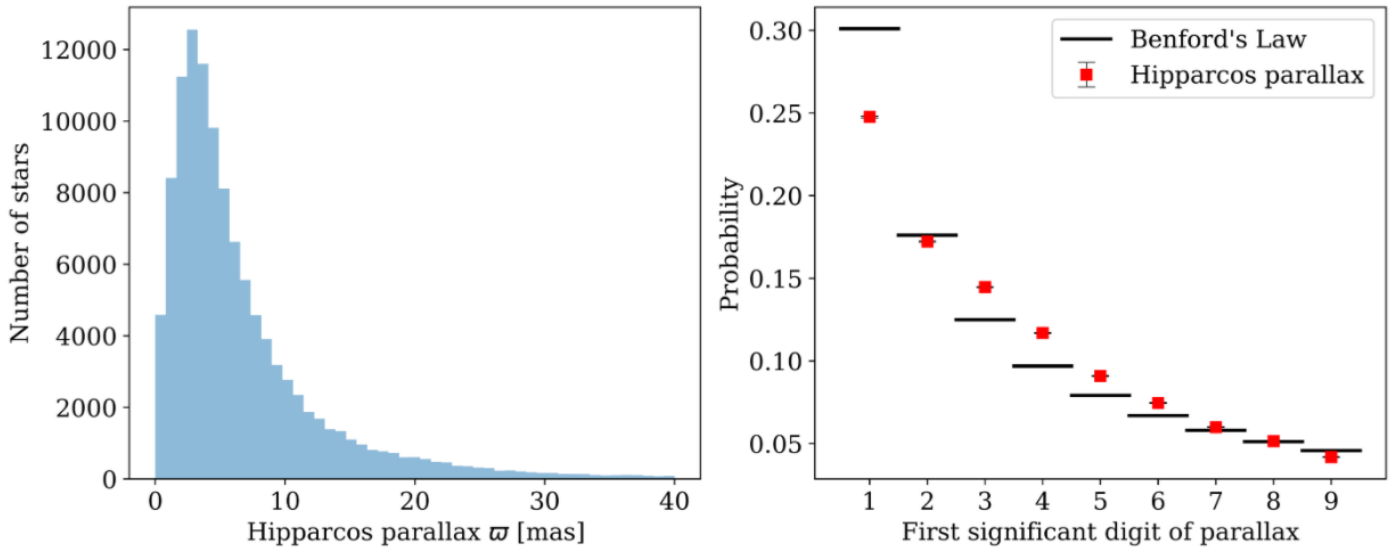

Fig. 3. *Left:* HIPPARCOS parallax histogram for all 113 942 stars from van Leeuwen (2007) with $\varpi > 0$ mas (1728 objects fall outside the plotted range). *Right:* distribution of the first significant digit of the HIPPARCOS parallaxes together with the theoretical prediction of Benford's law. The data have vertical error bars to reflect Poisson statistics, but the error bars are much smaller than the symbol sizes.

up to 14 kpc. Based on various tests we conducted, trying to reproduce the results presented by Alexopoulos & Leontsinis, we conclude the following:

– The assessment of Alexopoulos & Leontsinis has been based on the HYG 2.0 database (September 2011), which consists of the original HIPPARCOS catalogue (ESA 1997) that contains 118 218 entries, 117 955 of which have five-parameter astrometry (position, parallax, and proper motion), merged with the fifth edition of the Yale Bright Star Catalog that contains 9110 stars and the third edition of the Gliese Catalog of Nearby Stars that contains 3803 stars. We verified that changing the data set to the latest version of the catalogue, HYG 3.0 (November 2014), which is based on the new reduction of the HIPPARCOS data by van Leeuwen (2007), does not dramatically change the results.

– The assessment of Alexopoulos & Leontsinis has simply removed the small number of entries with non-positive parallax measurements. Presumably, this was done because negative parallaxes, which are a natural but possibly non-intuitive outcome of the astrometric measurement process underlying the HIPPARCOS and also the *Gaia* mission, cannot be directly translated into distance estimates (see the discussion in Sect. 4). Because the fraction of HIPPARCOS entries with zero or negative parallaxes is small (4245 and 4013 objects, corresponding to 3.6% and 3.4% for the data from ESA or van Leeuwen, respectively), excluding or including them does not fundamentally change the statistics.

– Distances have been estimated by Alexopoulos & Leontsinis from the parallax measurements through simple inversion. Whereas the true parallax and distance of a star are inversely proportional to each other, the estimation of a distance from a measured parallax, which has an associated uncertainty (and which can even be formally negative) requires care to avoid biases and to derive meaningful uncertainties (see e.g. Luri et al. 2018). While for relative parallax errors below ~10–20% a distance estimation by parallax inversion is an acceptable approach, Bayesian methods are superior for a distance estimation for larger relative parallax errors (see the discussion in Sect. 4.2). In the case of the HIPPARCOS data sets, only 42% and 51% of the objects from ESA and van Leeuwen, respectively, have relative parallax errors below 20%. In general, the "distances" inferred by Alexopoulos & Leontsinis are therefore biased as well as unreliable.

Figure 3 shows the histogram of the parallax measurements in the HIPPARCOS data set from van Leeuwen (2007) and the associated first-significant-digit distribution. The latter resembles Benford's law, at least trend-wise, although we recognise that it is statistically speaking not an acceptable description. Nonetheless, the overabundance of low-number compared to high-number digits is striking and significant.

Following Appendix F.3, the distribution of inverse parallaxes, that suggestively but incorrectly are referred to as "distances" by Alexopoulos & Leontsinis (2014), should also follow (the trends of) Benford's law. This is confirmed in Fig. 4. It shows that because small, positive parallaxes ($0 < \varpi \lesssim 1$ mas) are abundant in the HIPPARCOS data set, many stars are placed (well) beyond 1 kpc in inverse parallax ("distance"). As a result, the inverse parallaxes ("distances") span several orders of magnitude and resemble Benford's Law. We conclude that the results obtained by Alexopoulos & Leontsinis (2014) are reproducible but that their interpretation of inverse HIPPARCOS parallaxes as distances can be improved.

4. *Gaia* DR2

The recent release of the *Gaia* DR2 catalogue (Gaia Collaboration et al. 2018) offers a unique opportunity to make a study of Benford's law and stellar distances not based on 0.1 million, but 1000+ million objects. We discuss the parallaxes in Sect. 4.1 and the associated distance estimates in Sects. 4.2 and 4.3. We compare our findings with simulations in Sect. 4.4.

4.1. *Gaia* DR2 parallaxes

Gaia DR2 contains five-parameter astrometry (position, parallax, and proper motion) for 1 331 909 727 sources. One feature of the *Gaia* DR2 catalogue is the presence of spurious entries and non-reliable astrometry. Suspect data can be filtered out

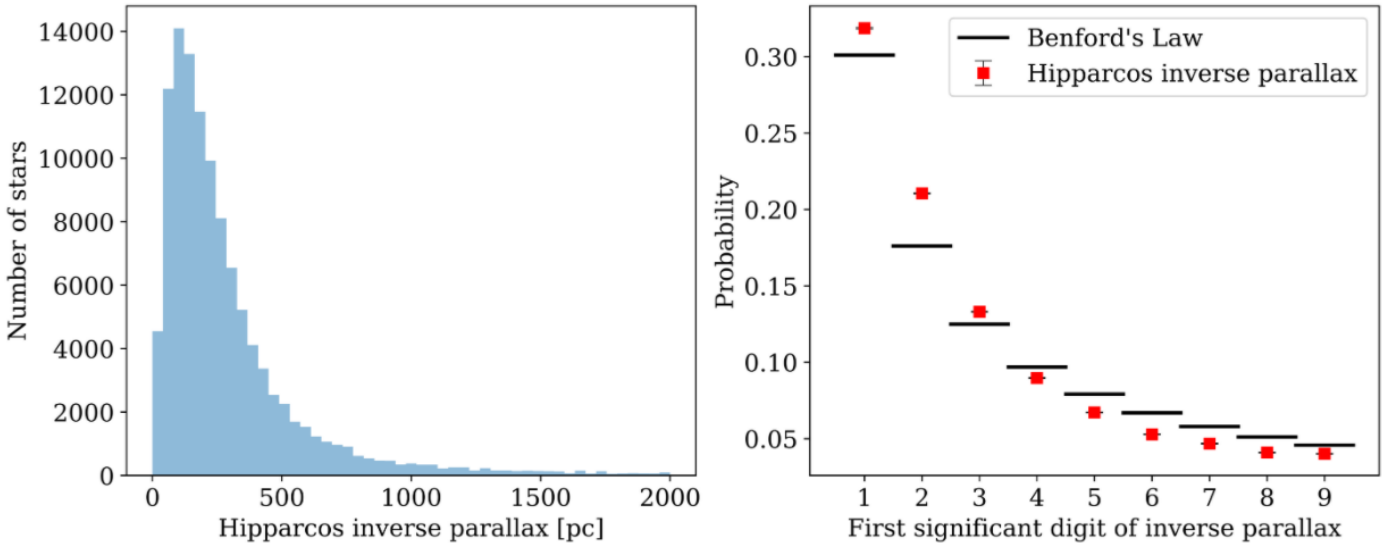

Fig. 4. *Left:* HIPPARCOS inverse-parallax histogram for all 113 942 stars from van Leeuwen (2007) with $\varpi > 0$ mas (3803 objects fall outside the plotted range). *Right:* distribution of the first significant digit of the inverse parallaxes, referred to as "distances" by Alexopoulos & Leontsinis (2014), together with the theoretical prediction of Benford's law; compare with Fig. 3b in Alexopoulos & Leontsinis (2014). The data have vertical error bars to reflect Poisson statistics, but the error bars are much smaller than the symbol sizes.

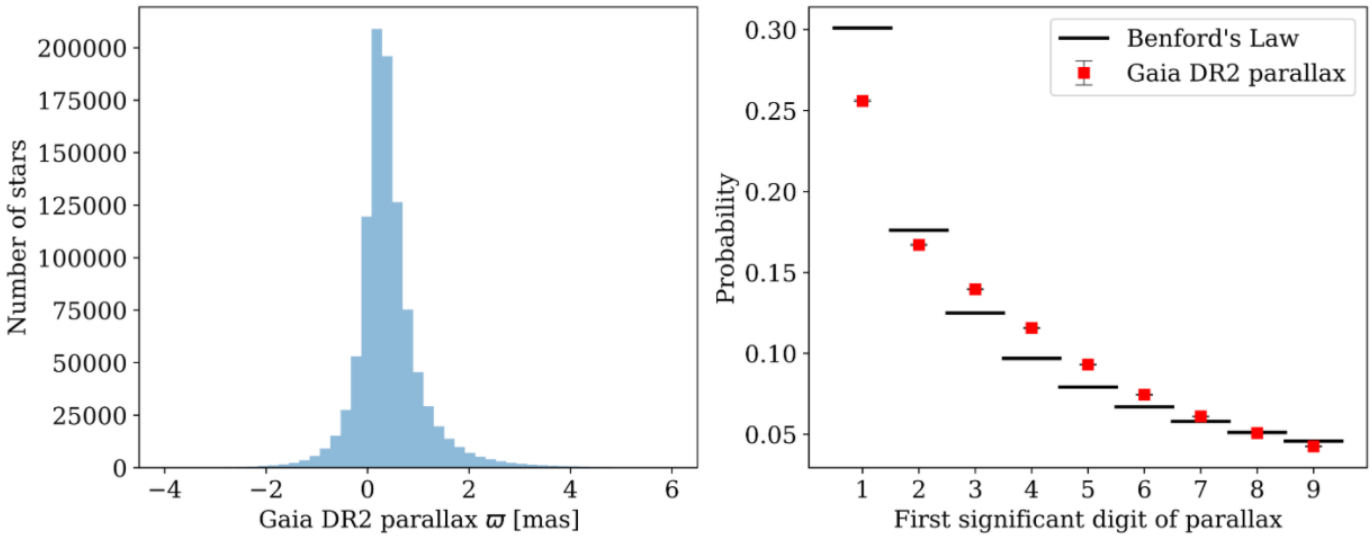

Fig. 5. *Left: Gaia* DR2 parallax histogram (for a random sample of one million objects); 2305 objects fall outside the plotted range. *Right:* distribution of the first significant digit of the absolute value of the *Gaia* DR2 parallaxes together with the theoretical prediction of Benford's law. The data have vertical error bars to reflect Poisson statistics, but the error bars are much smaller than the symbol sizes.

using quality filters recommended in Lindegren et al. (2018), Evans et al. (2018), and Arenou et al. (2018). Rather than filtering on the astrometric unit weight error (UWE), we filtered on the renormalised astrometric unit weight error (RUWE)[1]. In this study, we consistently applied the standard photometric excess factor filter published in April 2018 plus the revised astrometric quality filter published in August 2018 (see Appendix A for a detailed discussion). Another point worth mentioning is that in contrast to the HIPPARCOS case (Sect. 3), the percentage of non-positive parallaxes (*Gaia* DR2 does not contain objects with parallaxes that are exactly zero) in *Gaia* DR2 is substantial, at 26% (Fig. 5). Throughout this study, when we consider the first significant digit of *Gaia* DR2 parallaxes, we take the absolute value of the parallaxes first. The justification and implications of this choice are discussed in Appendix B.

Figure 5 shows the distribution of the first significant digit of the *Gaia* DR2 parallaxes. In practice, to save computational resources, we use randomly selected subsets of the data throughout this paper, unless stated otherwise (see Appendix C for a detailed discussion). The first significant digit of the parallax sample follows Benford's law well.

This is expected because at least two conditions identified in Sect. 2 are met. First of all, the histogram of the *Gaia* DR2 parallaxes shows that the distributions cover four orders of magnitude. Secondly, the distribution of the first significant digit of the parallaxes is not sensitive to scaling. That is, we verified that multiplying all parallaxes with a constant C does not fundamentally change the distribution of the first significant digit: for instance,

[1] See the *Gaia* DR2 known issues website `https://www.cosmos.esa.int/web/gaia/dr2-known-issues` and Appendix A for details.

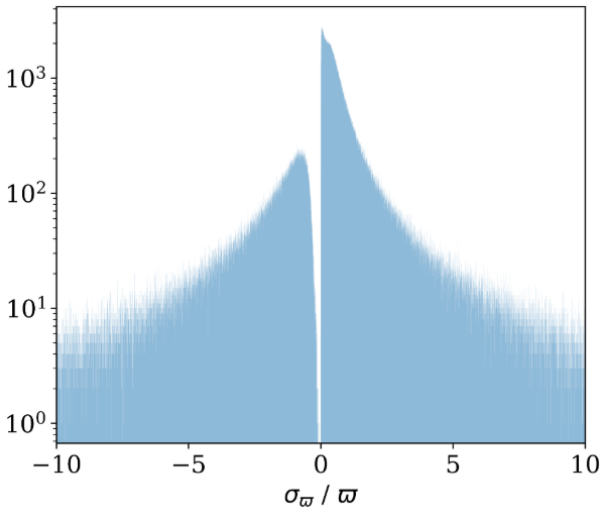

Fig. 6. Central part of the distribution of relative parallax errors in the *Gaia* DR2 catalogue (the inverse of field `parallax_over_error`).

the probabilities for digit $d = 1$ do not change by more than 5% points by scaling the data with any factor C in the range 1–10 (see also Appendix D). We postpone the discussion of why the first significant digit of the parallaxes follows Benford's law so closely to Sect. 4.4.

4.2. Gaia DR2 distance estimates

As we reported in Sect. 3, astrometry missions such as Hipparcos and *Gaia* do not measure stellar distances but parallaxes. These measurements are noisy such that as a result of the non-linear relation between (true) parallax and (true) distance (distance $\propto$ parallax^{-1}), distances estimated as inverse parallaxes are fundamentally biased (for a detailed disussion, see Luri et al. 2018). Whereas this bias is small and can hence be neglected, for small relative parallax errors (e.g. below ∼10–20%), it becomes significant for less precise data. Figure 6 shows a histogram of the relative parallax error of the *Gaia* DR2 catalogue. It shows that only 9.9% of the objects with positive parallax have $1/$`parallax_over_error` < 0.2, indicating that great care is needed in distance estimation.

As explained in Bailer-Jones et al. (2018, and references therein), distance estimation from measured parallaxes is a classical inference problem that is ideally amenable to a Bayesian interpretation. this approach has the advantage that negative parallax measurements can also be physically interpreted and that meaningful uncertainties on distance estimates can be reconstructed. Using a distance prior based on an exponentially decreasing space density (EDSD) model, Bailer-Jones et al. (2018) presented Bayesian distance estimates for (nearly) all sources in *Gaia* DR2 that have a parallax measurement. Figure 7 compares the distribution of the first significant digit of these distance estimates to Benford's law. This time, a poor match can be noted: instead of 1 as the most frequent digit, digits 2 and 3 appear more frequently. Why the first significant digits of the Bailer-Jones distance estimates do not follow Benford's law is evident from their histogram: Most stars in *Gaia* DR2 are located at ∼2–3 kpc from the Sun (see also Fig. 8). This is mostly explained by the EDSD prior adopted by Bailer-Jones et al. in their Bayesian framework. For the (small) set of (nearby) stars with highly significant parallax measurements, the choice of this prior is irrelevant and the distance estimates are strongly constrained by the measured parallaxes themselves. For the (vast) majority of (distant) stars, however, the low-quality parallax measurements contribute little weight, and the distance estimates mostly reflect the choice of the prior. The EDSD prior has one free parameter, namely the exponential length scale L, which can be tuned independently for each star. Bailer-Jones et al. (2018) opted to model this parameter as function of galactic coordinates (ℓ, b) based on a mock galaxy model. Because the EDSD prior has a single mode at $2L$ and because L (`r_len` in the data model) has been published along with the Bayesian distance estimates for each star, a prediction for the distribution of the first significant digit of the mode of the prior can be made accordingly. Figure 9 shows this prediction, along with the actual distribution of the mode of the EDSD prior, for a random sample of one million stars. The digit distribution compares qualitatively well with that of the Bayesian distance estimates (cf. Fig. 7), with digit 2 appearing most frequently, followed by digits 1 and 3, followed by digit 4, and with digits 5–9 being practically absent. Quantitative differences between the digit distributions can be understood by comparing the left panels of Figs. 7 and 9. Whereas the distance distribution has a smooth, Rayleigh-type shape, extending out to ∼8 kpc, the prior mode distribution is noisy as a result of the extinction law applied in the mock galaxy model used by Bailer-Jones et al. (2018) and lacks signal below ∼700 pc and above ∼5 kpc. This finite range is a direct consequence of the way in which the length scale was defined by Bailer-Jones et al., who opted to compute it for 49 152 pixels on the sky as one-third of the median of the (true) distances to all the stars from the galaxy model in that pixel (and subsequently creating a smooth representation as function of Galatic coordinates ℓ, b by fitting a spherical harmonic model). This resulted in a lowest value of L of 310 pc and a highest value of 3.143 kpc (such that the EDSD prior mode $2L$ can only take values between 620 pc and 6.286 kpc).

Anders et al. (2019) published a set of 265 637 087 photo-astrometric distance estimates obtained by combining *Gaia* DR2 parallaxes for stars with $G < 18$ mag with PanSTARRS-1, 2MASS, and AllWISE photometry based on the StarHorse code. The recommended quality filters SH_GAIAFLAG = "000" to select non-variable objects that meet the RUWE and photometric excess-factor filters from Appendix A and SH_OUTFLAG = "00000" to select high-quality StarHorse distance estimates leave 136 606 128 objects. Figure 10 shows their distance histogram and the associated distribution of the first significant digit. The strong preference for digits 1 and 2, followed by digits 7 and 8, is explained by the bi-modality of the distance histogram, showing a strong peak of main-sequence dwarfs at ∼1.5 kpc and a secondary peak of (sub)giants in the Bulge around 7.5 kpc.

Our main conclusion of this section is that all available, large-volume, *Gaia*-based distance estimates prefer small leading digits. This fact, however, foremost reflects the structure of the Milky Way, combined with its luminosity function and extinction law, and the magnitude-limited nature of the *Gaia* survey.

4.3. Gaia DR2 distance estimates in the solar neighbourhood

It is to be expected that the distribution of the first significant digit of stellar distances is (much) less dependent on the prior for high-quality parallaxes, for which the distance estimates are strongly constrained by the parallax measurements themselves. To verify this, we show this distribution in Fig. 11 for the subset of 243 291 stars in *Gaia* DR2 that have distance estimates below

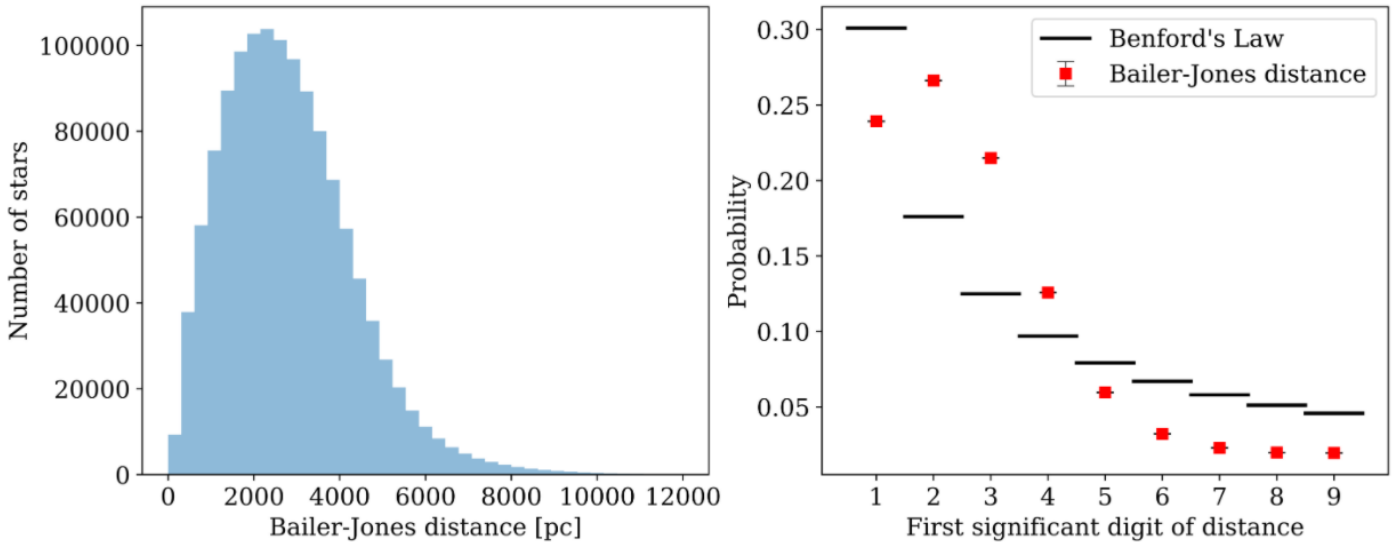

Fig. 7. *Left:* histogram of the *Gaia* DR2 Bayesian distance estimates from Bailer-Jones et al. (2018) for a random sample of one million objects (308 objects fall outside the plotted range). *Right*: distribution of the first significant digit of the Bayesian distance estimates, together with the theoretical prediction of Benford's law. The data have vertical error bars to reflect Poisson statistics, but the error bars are much smaller than the symbol sizes.

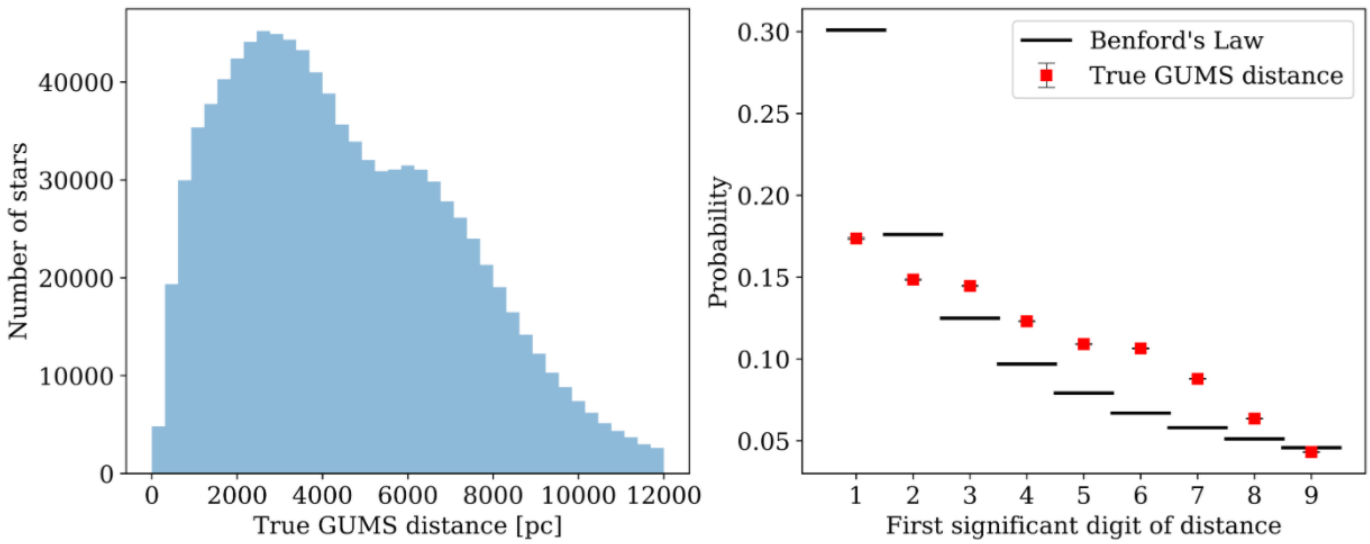

Fig. 8. *Left:* histogram of one million simulated true GUMS distances from Robin et al. (2012); 19 594 objects fall outside the plotted range. *Right*: distribution of their first significant digit together with the theoretical prediction of Benford's law. Figure 7 shows the same contents, but using the *Gaia* DR2 distance estimates from Bailer-Jones et al. (2018). The data have vertical error bars to reflect Poisson statistics, but the error bars are much smaller than the symbol sizes.

100 pc and hence have typically small relative parallax errors (e.g. the mean and median values of parallax over error are 261 and 214, respectively). In this case, in contrast to the case of Benford's law, digit 1 appears least frequently and digit 9 appears most frequently. Assuming that stars in the solar neighbourhood are approximately uniformly distributed with a constant density, this can be understood because the volume between equidistant (thin) shells centred on the Sun increases with the cube of the shell radius (e.g. there are $[100^3-90^3]/[20^3-10^3] \sim 39$ times as many objects in the shell between 10 and 20 pc compared to the shell between 90 and 100 pc). As shown in Fig. 11, the model of a constant-density solar neighbourhood is an almost perfect match with the data.

In order to determine out to which distance this is true, Figure 12 shows how the expectation value of the first significant digit of the distance distribution for all stars located within a sphere around the Sun with radius $R_\star$ varies with this radius and compares this expectation value with the one from the constant-density model. The (maybe initially surprising) variation in the expectation value between digits 2 and 7 with the distance limit of the sample (or the model) can be understood using the same argument as used before, linked to the cube dependence of the volume on distance. Fair agreement between data and model (difference <10%) occurs up to ~720 pc, which depending on extinction corresponds to early-M dwarfs for a faint limiting magnitude of 20.7 mag. To find a value of 720 pc is not surprising given the ~300 pc vertical scale height of the thin disc (e.g. Bland-Hawthorn & Gerhard 2016). We conclude that whereas the distribution of the first significant digit of distances of the full *Gaia* DR2 catalogue is biased to digits 2 and 3 (Sect. 4.2)

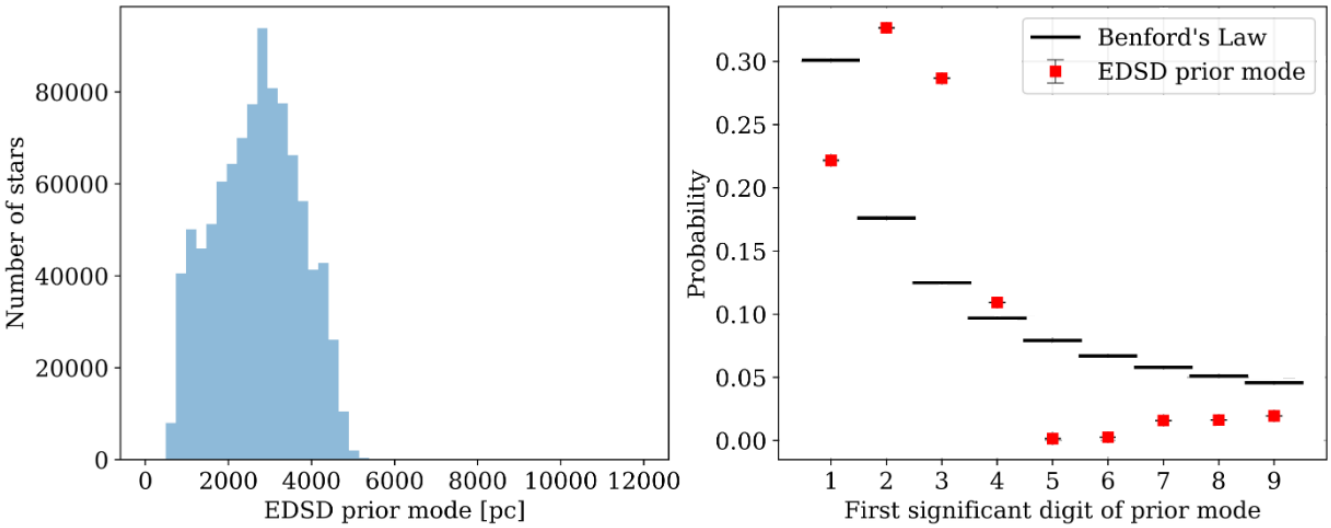

Fig. 9. *Left:* histogram of the mode of the EDSD prior (i.e. twice the exponential length scale $L = L(l, b)$) used by Bailer-Jones et al. (2018) for their Bayesian distance estimations of *Gaia* DR2 sources. *Right*: distribution of the first significant digit of the EDSD prior mode, together with the theoretical prediction of Benford's law. The data have vertical error bars to reflect Poisson statistics, but the error bars are much smaller than the symbol sizes.

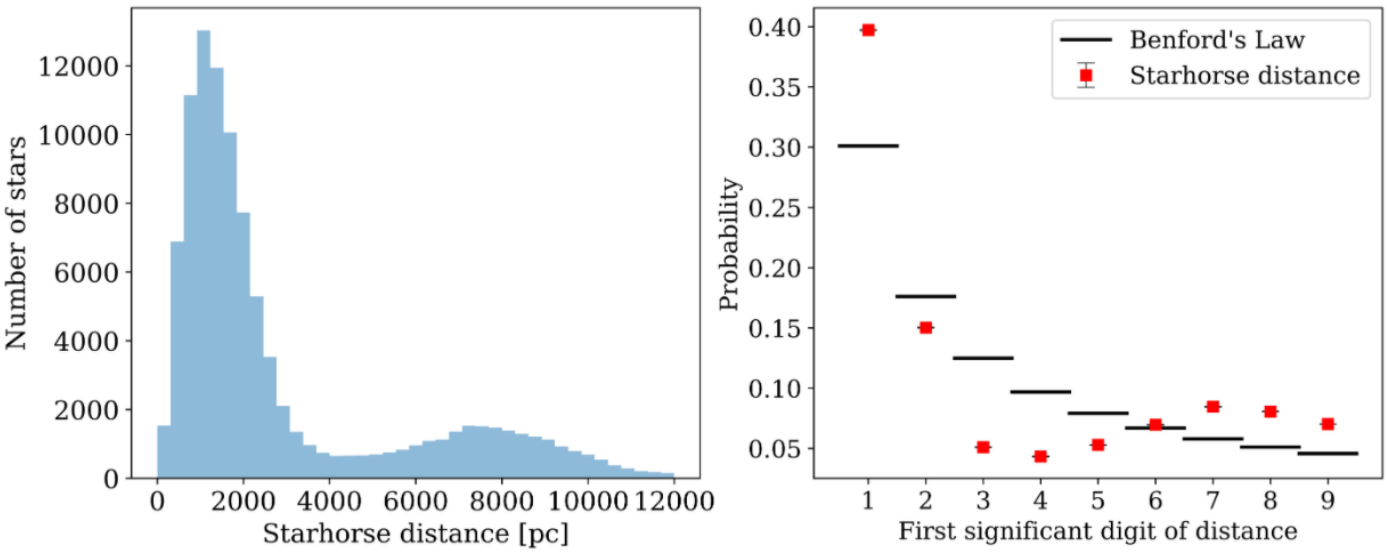

Fig. 10. *Left:* histogram of the subset of high-quality StarHorse distances from Anders et al. (2019). *Right*: distribution of the first significant digit of these distances, together with the theoretical prediction of Benford's law. The data have vertical error bars to reflect Poisson statistics, but the error bars are much smaller than the symbol sizes.

and mostly reflects the prior that has been used in the Bayesian estimation of the distances, local samples ($\lesssim$720 pc) with high-quality parallaxes show a preference for a range of digits, with the most frequent digit depending on the limiting distance, which is fully compatible with a distribution of stars with uniform, constant density.

4.4. Comparison with Gaia simulations

Robin et al. (2012) presented the *Gaia* universe model snapshot (GUMS). GUMS is a customised and extended incarnation of the Besançon galaxy model, fine-tuned to a perfect *Gaia* spacecraft that makes error-less observations. GUMS represents a sophisticated, realistic, simulated catalogue of the Milky Way (plus other objects accessible to *Gaia,* such as asteroids and external galaxies) observable by *Gaia*, containing more than one billion stars down to $G < 20$ mag. According to the Besançon galaxy model, the Milky Way consists of an exponentially thin disc (67% of the objects), an exponentially thick disc (22% of the objects), a bulge (10% of the objects), and a halo (1% of the objects). Not surprisingly, given the luminosity function and magnitude-limited nature of the *Gaia* survey, the majority of the stars in GUMS are within a few kiloparsec from the Sun, with 69% being main-sequence objects and 29% being sub(giants). Figure 8 shows the histogram of the true (i.e. noiseless, simulated) GUMS distances and the associated distribution of the first significant digits, which does not resemble Benford's Law at all. The aforementioned bi-modality in distances, with a main disc peak around 2–3 kpc and a strong secondary bulge peak around 5–7 kpc, explains the relatively even occurrence of the numbers 1–7 as leading significant digit.

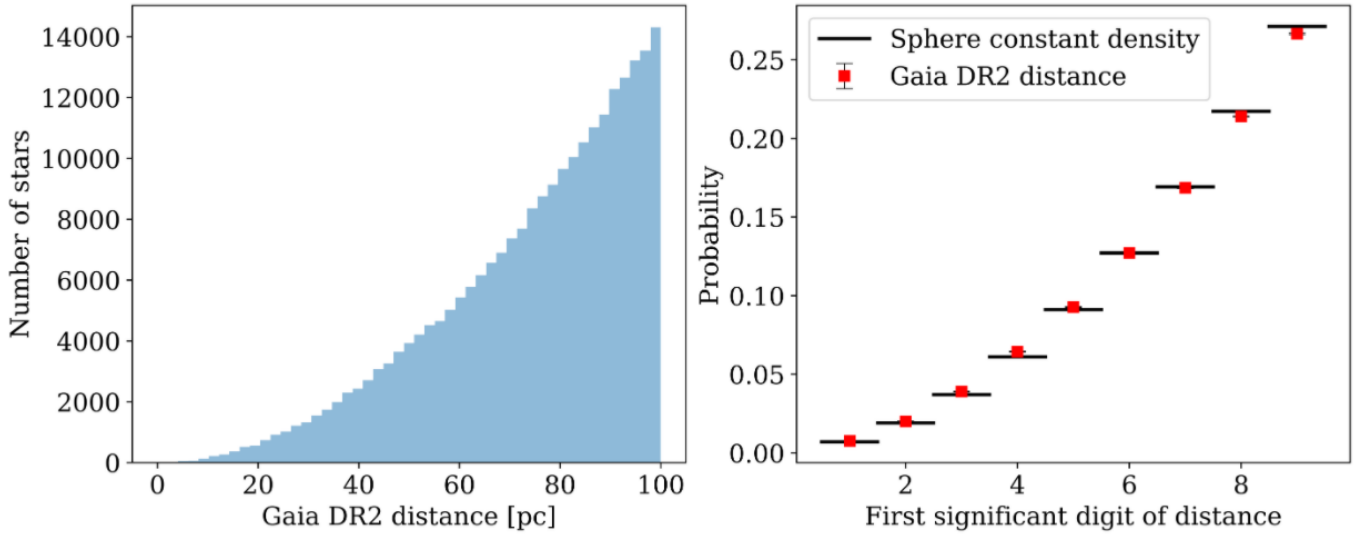

Fig. 11. *Left:* histogram of the *Gaia* DR2 Bayesian distance estimates from Bailer-Jones et al. (2018) for the sample of 243 291 stars within 100 pc from the Sun. *Right*: distribution of the first significant digit of the *Gaia* DR2 Bayesian distance estimates displayed in the left panel, together with the theoretical prediction of a sample of stars with uniform, constant density. The data have vertical error bars to reflect Poisson statistics, but the error bars are much smaller than the symbol sizes.

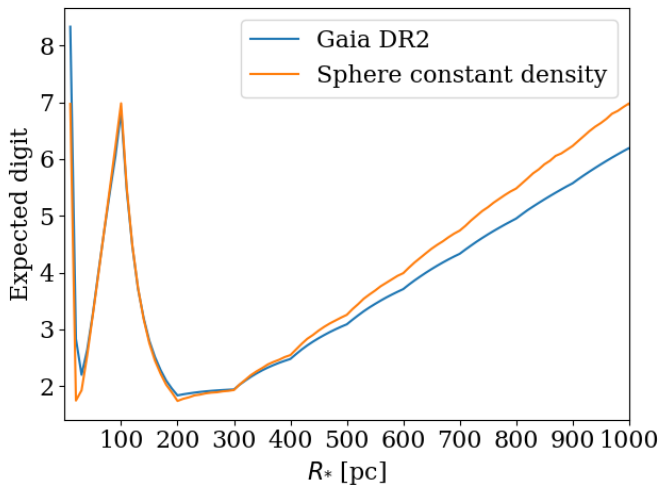

Fig. 12. Comparison of the expectation value of the first significant digit of the distance distribution of (a) all *Gaia* DR2 stars with distances less than $R_\star$ and (b) a model of the solar neighbourhood in which stars have a uniform, constant density.

Luri et al. (2014) presented an observed version of the GUMS catalogue resulting from application of *Gaia*-specific error models that implement realistic observational errors that depend, as in reality, on object properties such as magnitude, on the *Gaia* instrument characteristics such as read-out noise, and on the number of observations made over the nominal five-year operational lifetime. The vast majority of the 523 million individual, single stars are main-sequence dwarfs of spectral types F, G, K, and M (381 million, corresponding to 73%) and (sub)giants of spectral type F, G, and K (133 million, corresponding to 25%). Figure 13 shows the parallax distribution of this observed GUMS sample, together with the distribution of their first significant digit. When we compare Fig. 13 with Fig. 5, we recall that the first reflects a five-year *Gaia* mission and the second refers to *Gaia* DR2, which is based on 22 months of data (which implies that the formal uncertainties are to first order a factor $\sqrt{60/22} \approx 1.7$ larger). Nonetheless, the agreement between simulations and *Gaia* DR2 is striking.

When we compare Fig. 8 with Fig. 13, it is striking that the distribution of noise-free GUMS distances shows a larger departure from Benford's law than the distribution of noisy parallaxes simulated from them. When the noise-free GUMS distances are inverted to noise-free parallaxes, the Euclidean distance of the first-significant-digit distribution with respect to Benford's Law does not drastically change (<10%), suggesting that not the inversion in itself, but the observational (parallax) errors are responsible for the improved match to Benford's Law. This agrees with the inverse-invariance of first-significant-digit distributions that follow Benford's law (see Appendix F.3). In order to investigate this further, we took the noise-free GUMS parallaxes and perturbed them with a parallax error that was randomly drawn (for each individual object) from a Gaussian distribution with zero mean and a fixed standard deviation that reflects the parallax standard error. The first significant digits of the associated distribution of noisy parallaxes can then be compared to Benford's Law, and the agreement quantified through the Euclidean distance metric. When we repeated this for ever-increasing parallax standard errors (σ_ϖ), we found that the Euclidean distance with respect to Benford's Law rapidly decreased by a factor ~2, from 0.13 for $\sigma_\varpi = 0$ mas (noise-free GUMS parallaxes) to 0.06 for $\sigma_\varpi = 0.7$ mas (which is a typical parallax standard error for a representative faint star in *Gaia* DR2). In other words, the distribution of the first significant digits of the parallaxes approaches Benford's Law more and more when the parallaxes are made more and more noisy. The reason for this significant reduction in Euclidean distance when observational parallax errors are added is shown in Figs. 4 and 9 in Luri et al. (2014). These figures show that for the vast majority of stars, the (end-of-life, so surely the *Gaia* DR2) parallax error is larger than the true parallax itself. Combined with the fact that more than 40% of the GUMS stars have $d = 1$ as leading significant digit for their true parallax, as dictated by the nature of the Milky Way and the *Gaia* survey, this causes a bell-shape of the

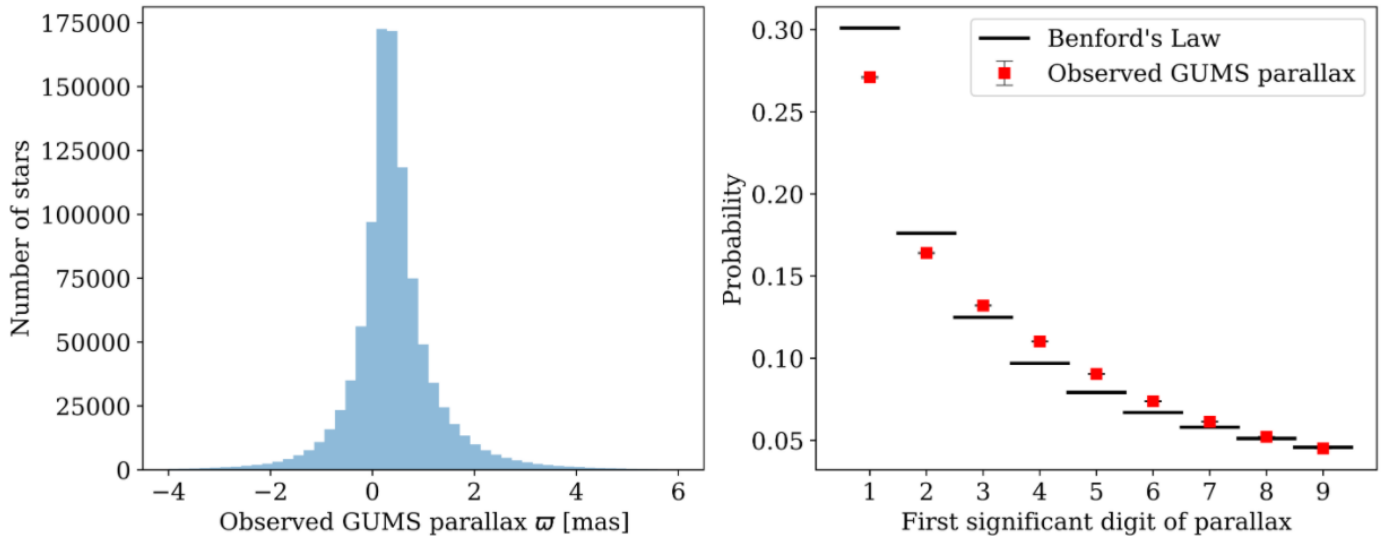

Fig. 13. *Left:* histogram of the simulated observed GUMS parallaxes from Luri et al. (2014); 4879 objects fall outside the plotted range. *Right:* distribution of their first significant digit together with the theoretical prediction of Benford's law. The data have vertical error bars to reflect Poisson statistics, but the error bars are much smaller than the symbol sizes. Figure 5 shows the same contents, but using the *Gaia* DR2 parallaxes.

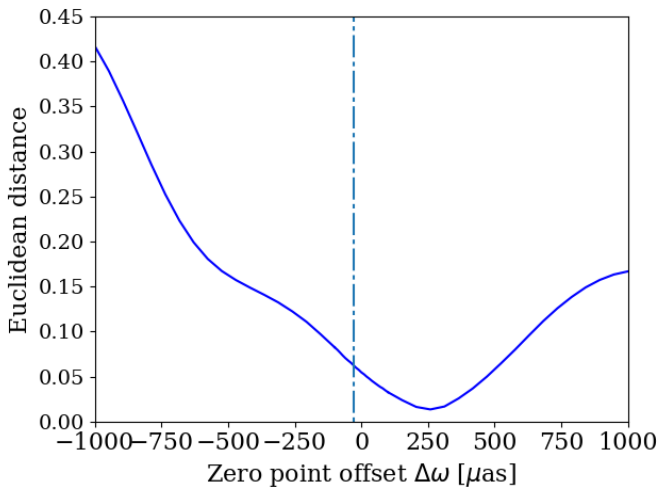

Fig. 14. Euclidean distance between the distribution of the first significant digit of the *Gaia* DR2 parallaxes (see Appendix D), after subtracting a trial zero-point offset $\Delta\varpi$ such that $\varpi_{\rm corrected} = \varpi_{Gaia}$ DR2 $- \Delta\varpi$, and Benford's law for trial zero-point offsets $\Delta\varpi$ between −1000 and 1000 μas. The dashed vertical line denotes the (faint-) QSO-based offset of −29 μas derived in Lindegren et al. (2018), with more recent work suggesting that the relevant value for (bright) stars is about −50 μas (see Sect. 5.1).

distribution of observed parallaxes around a low mean parallax value (Figs. 5 and 13, left panels) such that the first significant digits nicely follow Benford's Law (Figs. 5 and 13, right panels).

5. Parallax zero-point

5.1. Background

The global parallax zero-point offset in the *Gaia* DR2 data set should have come as no surprise. It has been known from HIPPARCOS times (e.g., Arenou et al. 1995; Makarov 1998) that a scanning, global space astrometry mission with a design such as that of HIPPARCOS and *Gaia* will have a (almost) full degeneracy between spin-synchronous variations of the basic angle between the (viewing directions of the) two telescopes and the zero-point of the parallaxes in the catalogue (for details, see Butkevich et al. 2017). The zero-point offset was determined during the data processing and was published in Lindegren et al. (2018) as −29 ± 1 μas, in the sense of *Gaia* parallaxes being too small, based on the median parallax of a sample of half a million primarily faint quasars contained in *Gaia* DR2. During the internal validation of the data processing prior to release, the zero-point was investigated using ~30 different methods and samples, systematically resulting in a negative offset of order a few dozen μas (see Table 1 in Arenou et al. 2018). During these early inspections, hints already appeared that the zero-point offset depends on sky position, magnitude, and colour of the source.

Subsequent external studies, using a variety of methods and primarily bright stellar samples, and often combined with external data, often resulted in a zero-point offset of about −50 μas. Some recent examples, ordered from low to high offset values, include −28 ± 2 μas from Shao & Li (2019) based on mixture modelling of globular clusters, −31 ± 11 μas from Graczyk et al. (2019) based on eclipsing binaries, −35 ± 16 μas from Sahlholdt & Silva Aguirre (2018) based on asteroseismology, −41 ± 10 μas from Hall et al. (2019) based on asteroseismology, −42 ± 13 μas from Layden et al. (2019) based on RR Lyraes, −46 ± 13 μas from Riess et al. (2018) based on classical Cepheids, −48 ± 1 μas from Chan & Bovy (2020) based on hierarchical modelling of red clump stars, −49 ± 18 μas from Groenewegen (2018) based on classical Cepheids, −50 ± 5 μas from Khan et al. (2019) based on asteroseismology, −52 ± 2 μas from Leung & Bovy (2019) based on APOGEE spectrophotometric distances, −53 ± 9 μas from Zinn et al. (2019) based on asteroseismology and spectroscopy, −54 ± 6 μas from Schönrich et al. (2019) based on *Gaia* DR2 radial velocities, −57 ± 3 μas from Muraveva et al. (2018) based on RR Lyraes, −75 ± 29 μas from Xu et al. (2019) based on VLBI astrometry, −76 ± 25 μas from Lindegren (2020) based on VLBI data of radio stars, and −82 ± 33 μas from Stassun & Torres (2018) based on eclipsing binaries.

Table 1. Frequency of occurrence of the first significant digit of the *Gaia* DR2 parallaxes (first line), a Lorentzian distribution with half-width $\gamma = 360\,\mu$as (second line), a Lorentzian distribution with half-width $\gamma = 360\,\mu$as and shifted by $+260\,\mu$as (third line), and Benford's law (fourth line).

1	2	3	4	5	6	7	8	9	Data/function
0.256	0.167	0.140	0.116	0.093	0.075	0.061	0.051	0.043	*Gaia* DR2 parallaxes
0.289	0.181	0.132	0.102	0.081	0.067	0.056	0.049	0.044	Lorentzian
0.260	0.174	0.140	0.112	0.090	0.072	0.059	0.049	0.043	Lorentzian shifted by $+260\,\mu$as
0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046	Benford's law

5.2. Varying the parallax zero-point offset

The question arises whether the already fair agreement between the *Gaia* DR2 parallaxes and Benford's law, as discussed in Sect. 4.1 and displayed in Fig. 5, would further improve when due account of the parallax zero-point offset would be taken. Naively, we would expect that a distribution of a quantity such as the parallax that covers several orders of magnitude, a small, uniform shift would not drastically change its behaviour with respect to Benford's law.

Our findings are summarised in Fig. 14. It shows for a range of trial zero-point offsets $\Delta\varpi$ the Euclidean distance between the distribution of the first significant digit of the *Gaia* DR2 parallaxes, after subtracting a zero-point offset $\Delta\varpi$ such that $\varpi_{\rm corrected} = \varpi_{Gaia\ {\rm DR2}} - \Delta\varpi$, and Benford's law. With this convention, the zero-point offset from Lindegren et al. (2018) translates into $\Delta\varpi = -29\,\mu$as (see Sect. 5.1). The plot shows that changing the offset from $\Delta\varpi = 0$ to $-29\,\mu$as only changes the Euclidean distance metric by 0.007, and even in the direction of worsening the agreement between the offset-corrected parallaxes and Benford's law.

A striking feature in Fig. 14 is the pronounced minimum seen around $\Delta\varpi \sim +260\ \mu$as. This minimum can be understood as follows. The *Gaia* DR2 parallax histogram itself (Fig. 5) roughly resembles a Lorentzian of half-width $\gamma = 360\,\mu$as and with a mean that is offset by some $-260\,\mu$as. Table 1 shows that such a Lorentzian has a distribution of first significant digits that already resembles that of Benford's law (see Appendix F.4), with digit 1 appearing most frequently and digit 9 appearing least frequently. By applying a uniform offset of $+260\ \mu$as, the corrected parallax distribution becomes roughly symmetric and the match between the Lorentzian and the shifted *Gaia* DR2 data improves even further. We conclude that the conspicuous minimum in Fig. 14 around $\Delta\varpi \sim +260\,\mu$as has a mathematical reason, namely that this particular parallax offset causes the distribution of the shifted *Gaia* DR2 parallaxes to become optimally symmetric, instead of being caused by the zero-point offset.

6. Summary and conclusions

We investigated whether Benford's law applies to *Gaia* DR2 data. Although it has been known for a long time that this law applies to a wide variety of physical data sets, it was only recently shown by Alexopoulos & Leontsinis (2014) that it also holds for HIPPARCOS astrometry. We showed that the 1.3 billion observed parallaxes in *Gaia* DR2 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 digit 9.

We reached a very different conclusion concerning the astrometric distance estimates. Using HIPPARCOS astrometry, Alexopoulos & Leontsinis (2014) computed distance estimates as the reciprocal of the parallax, and found that this data set also follows Benford's law closely. However, Bailer-Jones (2015) and Luri et al. (2018) showed that the reciprocal of the observed parallax ϖ^{-1} is a poor estimate of the distance when the relative parallax error exceeds ~10–20%. The distance estimate can be improved by adding prior information about our Galaxy (Bailer-Jones 2015) and/or by including additional data such as photometry (Anders et al. 2019). We unambiguously demonstrated that in neither case does the improved distance estimates follow Benford's law, although distances with small starting digits are still more abundant. Moreover, using realistic simulations of the stellar content of the Milky Way (Robin et al. 2012), we showed that the distances ought not to follow Benford's law, essentially because the interplay between the luminosity function of the Milky Way and *Gaia* mission selection function results in a bimodal distance distribution, corresponding to nearby dwarfs in the Galactic disc and distant giants in the Galactic bulge. The fact that the true distances underlying the *Gaia* catalogue do not follow Benford's law, while the observed parallaxes do follow this law, probably due to observational errors, is the most intriguing result of this paper.

One of our objectives was to use Benford's law (or the deviation from it) as an indicator of anomalous behaviour, not necessarily giving hard evidence, but rather providing an indicator whose subsets warrant a deeper analysis (e.g. Badal-Valero et al. 2018, investigating money laundering). We investigated the application of several astrometric and photometric quality filters applied to the *Gaia* DR2 parallaxes, but none changed the adherence to Benford's law by more than a few percent points.

Finally, we analysed the parallax zero-point that would be needed to optimise the fit to Benford's law, to compare it with the roughly $-50\,\mu$as zero-point offset that is known to be present in the *Gaia* DR2 parallaxes (e.g. Khan et al. 2019). An offset value of $+260\,\mu$as was recovered. This can be understood by the negative tail of the Lorentzian-like *Gaia* DR2 parallax distribution, which for this offset value results in an optimally symmetric corrected-parallax distribution that closely follows Benford's law. We therefore conclude that Benford's law should not be used to validate the parallax zero-point in *Gaia* DR2.

Acknowledgements. 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. We would like to thank Timo Prusti, Daniel Michalik, Alice Zocchi, and Eero Vaher for stimulating discussions and Eduard Masana for tips on how to create a (pseudo-)random query on the Paris-Meudon TAP server containing observed GUMS parallaxes. We would like to thank the anonymous referee for her/his constructive feedback. This research is based on data obtained by ESA's HIPPARCOS satellite and has made use of the ADS (NASA), SIMBAD/VizieR (CDS), and TOPCAT (Taylor 2005, http://www.starlink.ac.uk/topcat/) tools. We acknowledge support from the ESTEC Faculty Visiting Scientist Programme. The research leading to these results has received funding from the BELgian federal Science Policy Office (BELSPO) through PRODEX grants *Gaia* and *PLATO*.

References

Alexopoulos, T., & Leontsinis, S. 2014, J. Astrophys. Astron., 35, 639
Anders, F., Khalatyan, A., Chiappini, C., et al. 2019, A&A, 628, A94
Arenou, F., Lindegren, L., Froeschle, M., et al. 1995, A&A, 304, 52
Arenou, F., Luri, X., Babusiaux, C., et al. 2018, A&A, 616, A17
Badal-Valero, E., Alvarez-Jareño, J. A., & Pavía, J. M. 2018, Forensic Sci. Int., 22,
Bailer-Jones, C. A. L. 2015, PASP, 127, 994
Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Mantelet, G., & Andrae, R. 2018, AJ, 156, 58
Benford, F. 1938, Proc. Amer. Philos. Soc., 78, 551
Berger, A., & Hill, T. P. 2015, An Introduction to Benford's Law (Princeton University Press)
Bland-Hawthorn, J., & Gerhard, O. 2016, ARA&A, 54, 529
Butkevich, A. G., Klioner, S. A., Lindegren, L., Hobbs, D., & van Leeuwen, F. 2017, A&A, 603, A45
Chan, V. C., & Bovy, J. 2020, MNRAS, 493, 4367
ESA, 1997, in The HIPPARCOS and TYCHO Catalogues. Astrometric and Photometric Star Catalogues Derived from the ESA HIPPARCOS Space Astrometry Mission, ESA SP, 1200
Evans, D. W., Riello, M., De Angeli, F., et al. 2018, A&A, 616, A4
Gaia Collaboration (Brown, A. G. A., et al.) 2018, A&A, 616, A1
Gaia Collaboration (Prusti, T., et al.) 2016, A&A, 595, A1
Goodman, W. M. 2016, Significance, 13, 38
Graczyk, D., Pietrzyński, G., Gieren, W., et al. 2019, ApJ, 872, 85
Groenewegen, M. A. T. 2018, A&A, 619, A8
Hall, O. J., Davies, G. R., Elsworth, Y. P., et al. 2019, MNRAS, 486, 3569
Hill, T. P. 1995a, Stat. Sci., 10, 354
Hill, T. P. 1995b, Proc. Amer. Math. Soc., 123, 887
Khan, S., Miglio, A., Mosser, B., et al. 2019, The Gaia Universe, 13
Layden, A. C., Tiede, G. P., Chaboyer, B., Bunner, C., & Smitka, M. T. 2019, AJ, 158, 105
Lesperance, M., Reed, W. J., Stephens, M. A., Tsao, C., & Wilton, B. 2016, PLoS One, 11, e0151235
Leung, H. W., & Bovy, J. 2019, MNRAS, 489, 2079
Lindegren, L. 2020, A&A, 637, C5
Lindegren, L., Hernández, J., Bombrun, A., et al. 2018, A&A, 616, A2
Luri, X., Palmer, M., Arenou, F., et al. 2014, A&A, 566, A119
Luri, X., Brown, A. G. A., Sarro, L. M., et al. 2018, A&A, 616, A9
Makarov, V. V. 1998, A&A, 340, 309
Muraveva, T., Delgado, H. E., Clementini, G., Sarro, L. M., & Garofalo, A. 2018, MNRAS, 481, 1195
Newcomb, S. 1881, Am. J. Math., 4
Nigrini, M. J. 1996, J. Am. Taxation Assoc.: Publ. Tax Sect. Am. Acc. Assoc., 18, 72
Ochsenbein, F., Bauer, P., & Marcout, J. 2000, A&AS, 143, 23
Riess, A. G., Casertano, S., Yuan, W., et al. 2018, ApJ, 861, 126
Robin, A. C., Luri, X., Reylé, C., et al. 2012, A&A, 543, A100
Sahlholdt, C. L., & Silva Aguirre, V. 2018, MNRAS, 481, L125
Schönrich, R., McMillan, P., & Eyer, L. 2019, MNRAS, 487, 3568
Shao, Z., & Li, L. 2019, MNRAS, 2241,
Stassun, K. G., & Torres, G. 2018, ApJ, 862, 61
Tam Cho, W. K., & Gaines, B. J. 2007, Amer. Statist., 61, 218
Taylor, M. B. 2005, in Astronomical Data Analysis Software and Systems XIV, eds. P. Shopbell, M. Britton, & R. Ebert, ASP Conf. Ser., 347, 29
van Leeuwen, F. 2007, in Hipparcos, the New Reduction of the Raw Data, Astrophys. Space Sci. Lib., 350
Weisstein, E. W. 2019, MathWorld - A Wolfram Web Resource, http://mathworld.wolfram.com/BenfordsLaw.html
Xu, S., Zhang, B., Reid, M. J., Zheng, X., & Wang, G. 2019, ApJ, 875, 114
Zinn, J. C., Pinsonneault, M. H., Huber, D., & Stello, D. 2019, ApJ, 878, 136

Appendix A: Effect of *Gaia* DR2 quality filters

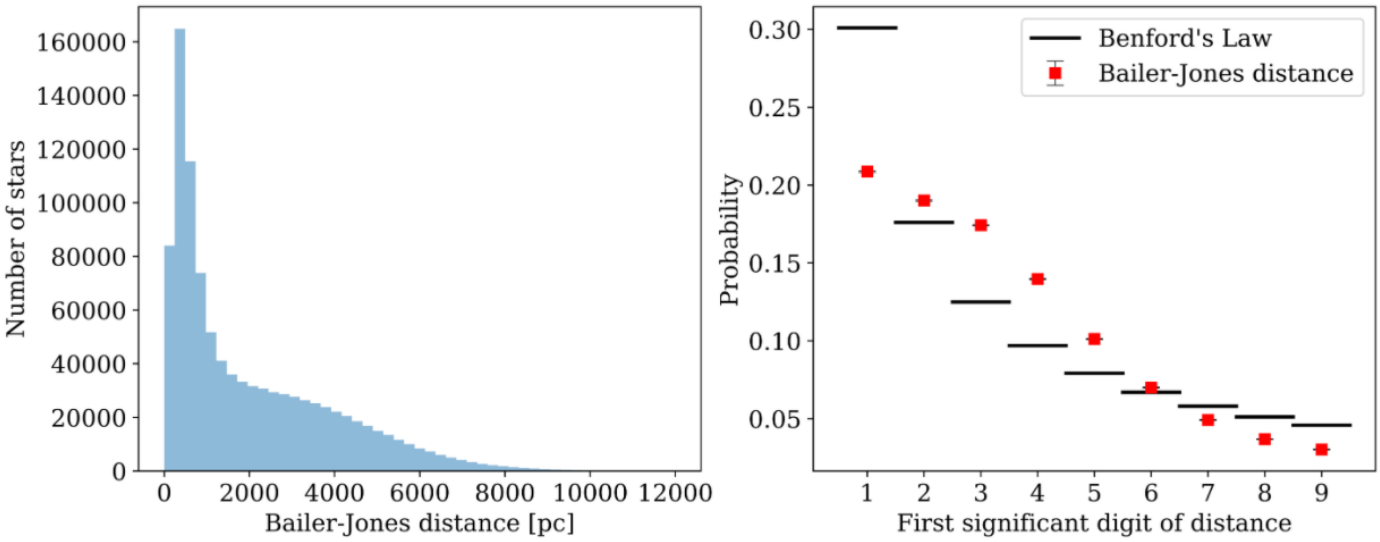

Fig. A.1. *Left:* histogram of the *Gaia* DR2 Bayesian distance estimates from Bailer-Jones et al. (2018) for the sample of one million objects with the highest RUWE values, i.e. the objects with the poorest astrometric quality. *Right*: distribution of the first significant digit of the Bayesian distance estimates, together with the theoretical prediction of Benford's law. The data have vertical error bars to reflect Poisson statistics, but the error bars are much smaller than the symbol sizes. Compare with Fig. 7 for a sample of one million random stars that meet all astrometric (and photometric) quality criteria.

Lindegren et al. (2018), Evans et al. (2018), and Arenou et al. (2018), all of whom were published together with and at the same date as the *Gaia* DR2 catalogue (25 April 2018), advocated using quality filters to define clean *Gaia* DR2 samples that are not hindered by astrometric and/or photometric artefacts. Such artefacts are known to be present in the data in particular in dense regions, and reflect the iterative and non-final nature of the data-processing strategy and status underlying *Gaia* DR2 and can be linked to erroneous observation-to-source matches, background subtraction errors, uncorrected source blends, etc. In this study, we employed the photometric excess factor filter as well as the astrometric quality filter, which is based on the renormalised unit weight error (RUWE) published post *Gaia* DR2 (in August 2018), requiring that valid sources meet the following two conditions:

$$1.0 + 0.015C^2 < E < 1.3 + 0.06C^2, \tag{A.1}$$

and

$$\mathrm{RUWE} = \frac{\sqrt{\chi^2/(N-5)}}{u_0(G,C)} < 1.4, \tag{A.2}$$

where in the notation of the *Gaia* DR2 data model, E = `phot_bp_rp_excess_factor` = (`phot_bp_mean_flux` + `phot_rp_mean_flux`)/`phot_g_mean_flux`, C = `bp_rp` = `phot_bp_mean_mag` − `phot_rp_mean_mag`, χ^2 = `astrometric_chi2_al`, N = `astrometric_n_good_obs_al`, G = `phot_g_mean_mag`, and $u_0(G,C)$ is a look-up table as function of G magnitude and $BP - RP$ colour index that is provided on the *Gaia* DR2 known issues webpage[2]. These filters combined remove 620 842 302 entries from the *Gaia* DR2 catalogue (corresponding to 47% of the data). The distribution of the first significant digit of the parallaxes is affected by application of the filters, but not to the extent that overall trends change. The maximum difference occurs for the frequency of digit 1, which equals 0.28 without filtering and 0.26 with the filtering applied.

Interestingly, and as a side note, the Bayesian distance estimates and associated first significant distribution of the sample of one million objects with the poorest astrometric quality (i.e. the highest RUWE values), displayed in Fig. A.1, differ substantially from those derived from a random sample of filtered stars, as displayed in Fig. 7. With the evidence provided in the *Gaia* DR2 documentation and in the references quoted above that the astrometric quality filter is effective in removing genuinely bad and suspect entries, this is no surprise.

Appendix B: Negative parallaxes

Figure B.1 shows that a significant fraction of the *Gaia* DR2 parallaxes is negative (26% of published data, which reduces to 18% when the filters discussed in Appendix A are applied). The complication of this fact is that there is no rule of how to deal with these data in relation to Benford's law. In addition, it is known that negative parallaxes are a totally normal and expected outcome of the astrometric measurement concept of *Gaia* and that simply ignoring negative parallaxes can lead to severe biases in the astrophysical interpretation of the data (for a detailed discussion, see Luri et al. 2018). Figure B.1 shows how sensitive the frequency distribution of the first significant digits is to the way in which we treat negative parallaxes: Variations in the frequency of occurrence of a few percent points appear depending on whether we include the negative parallaxes (after taking the absolute value) or not. Throughout this work, when we studied *Gaia* DR2 parallaxes and first-significant-digit statistics, we included all parallaxes, that is, we considered the absolute value of the measured parallax. This refers to the red symbols in Fig. B.1, which can be interpreted as a reasonable (weighted-average) compromise between the two limiting cases defined by objects with positive parallax on the one hand (green symbols)

[2] https://www.cosmos.esa.int/web/gaia/dr2-known-issues

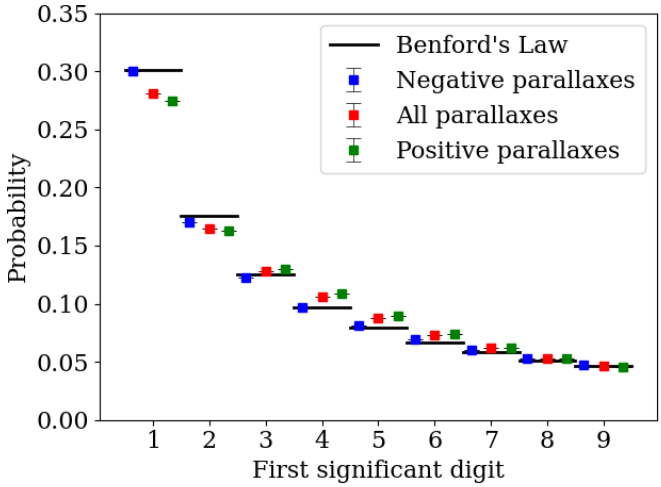

Fig. B.1. Distribution of the first significant digit of the *Gaia* DR2 parallaxes together with the theoretical prediction of Benford's law. Blue points refer to negative parallaxes (for which the statistics is based on the absolute value of the parallaxes), green points refer to positive parallaxes, and red points refer to the absolute value of all parallaxes (the red data can hence be considered as a weighted mean of the blue and green data with weights 18% and 82%, respectively). The data have vertical error bars to reflect Poisson statistics, but the error bars are much smaller than the symbol sizes.

and objects with negative parallax on the other hand (blue symbols).

Appendix C: Tests with smaller sample sizes

In view of the significant number of objects contained in *Gaia* DR2 and the associated non-negligible processing loads and run times, we conducted experiments to verify to what extent reduced sample sizes with randomly[3] selected objects return reliable results on the frequency distribution of the first significant digit of the parallaxes. Table C.1 summarises our findings. It shows that percent-level accurate data can be derived from randomly selected samples of about one million objects. When we refer to *Gaia* DR2 statistics here, we consistently used samples containing one million randomly selected entries without inducing loss of generality.

Appendix D: Statistics and Benford's law: justification of the Euclidean distance measure

We used the Euclidean distance to quantify how well the distribution of the first significant digit of *Gaia* data sets is described by Benford's law,

$$\mathrm{ED} = \sqrt{\Sigma_{d=1}^{9} (p_d - e_d)^2}, \tag{D.1}$$

where p_d is the measured digit frequency and e_d is the expected digit frequency for digit d according to Benford's law. The Euclidean distance ranges between 0 (when all first significant digits exactly follow Benford's law) and 1.036 (when all first significant digits equal 9), with lower values indicating better adherence to Benford's Law. Although the Euclidean distance is not a formal test statistic (see the discussion below), it is independent of sample size, which makes it suitable as a relative metric. The problem with sample-size dependent tests such as χ^2 (or Kolmogorov-Smirnov) is evident from their definition,

$$\chi^2 = \sum_{d=1}^{9} \left(\frac{O_d - E_d}{\sigma} \right)^2, \tag{D.2}$$

where O_d is the observed number of occurrences of digit d, E_d is the expected number of occurrences of digit d, and σ reflects the measurement error on O_d,

$$\begin{aligned} O_d &= N p_d \\ E_d &= N e_d = N \log_{10}\left(1 + \frac{1}{d}\right), \end{aligned} \tag{D.3}$$

where N is the total number of data points (stars in *Gaia* DR2 in our case, so $N \sim 10^9$). With counting (Poisson) statistics, $\sigma \propto \sqrt{N}$, such that $\chi^2 \propto N$. In practice, this means that a formal test statistic based on χ^2 would reject many cases in which the data under test come from distributions that follow Benford's law (e.g. Tam Cho & Gaines 2007). In other words, because χ^2 tests have large statistical power for high values of N, even quite small differences will be statistically significant, and because *Gaia* DR2 contains a billion parallaxes that are distributed over just nine bins ($d = 1, \ldots, 9$), the formal error bars on the observed frequencies in each bin (resulting just from Poisson statistics) are so tiny that any peculiarity in the data (e.g. an open cluster at

Table C.1. Frequency of occurrence of the first significant digit of the parallaxes in *Gaia* DR2 using randomly selected entries for various sample sizes.

First significant digit	1K	10K	100K	1M	10M	100M	*Gaia* DR2
1	0.292	0.284	0.281	0.281	0.281	0.281	0.281
2	0.160	0.159	0.167	0.164	0.164	0.164	0.164
3	0.122	0.125	0.127	0.128	0.128	0.128	0.128
4	0.098	0.109	0.105	0.106	0.106	0.106	0.106
5	0.106	0.0877	0.0870	0.0874	0.0876	0.0876	0.0876
6	0.0740	0.0701	0.0734	0.0728	0.0729	0.0729	0.0729
7	0.0671	0.0627	0.0620	0.0616	0.0617	0.0617	0.0617
8	0.0468	0.0548	0.0514	0.0529	0.0528	0.0528	0.0528
9	0.0356	0.0496	0.0463	0.0465	0.0460	0.0460	0.0460

Notes. The column *Gaia* DR2 refers to 1332M. K stands for 1000, while M stands for 1 000 000.

[3] We used the `random_index` field in *Gaia* DR2; see https://gea.esac.esa.int/archive/documentation/GDR2/Gaia_archive/chap_datamodel/sec_dm_main_tables/ssec_dm_gaia_source.html

a specific distance or an extragalactic objects), would affect the statistical test and might provide misleading conclusions (exclusively looking at statistical errors, compliance of *Gaia* DR2 parallaxes with Benford's Law would be excluded right away with very high significance levels). A reduced-χ^2 statistic would also not alleviate this problem because the reduction would only divide χ^2 by the number of bins (nine) and not by the number of data points (N). The Euclidean distance employed in this work, on the other hand, is independent of sample size and hence provides a metric that only becomes more precise with increasing sample size, but does not run away.

Clearly, a disadvantage of using the Euclidean distance is that it is not a formal test statistic with associated statistical power (although Goodman (2016) suggested that data can be said to follow Benford's law when the Euclidean distance is shorter than ~0.25). Many researchers have investigated and have proposed suitable metrics that can quantify statistical (dis)agreement between data and Benford's law (e.g. the Cramér-von Mises metric; Lesperance et al. 2016). We did not explore such metrics further for several reasons that are essentially all linked to the existing freedom and arbitrariness in the interpretation of the *Gaia* DR2 data, as listed below. Note here that we mean that a 5% increase of 0.4 (40%) results in 0.42 (42%), while a 5% point increase of 0.4 (40%) results in 0.45 (45%).

1. It is known that *Gaia* DR2 contains in addition to a small fraction of non-filtered duplicate sources (cf. Fig. 2 in Arenou et al. 2018) genuine sources with spurious astrometry (and/or photometry). As explained in Appendix A, several filters have been recommended to obtain clean data sets (e.g. RUWE, astrometric excess noise, photometric excess factor, number of visibility periods, and the longest semi-major axis of the five-dimensional astrometric error ellipsoid; Lindegren et al. 2018). In all of these cases, however, the specific threshold values to be used, and also which combination of filters to be used, is specific to the science application, without an absolute truth. Depending on subjective choices, the observed distribution over the first significant digits changes by up to several percent points (see Appendix A), which is orders of magnitude larger than the formal statistical errors (one billion stars divided equally over nine bins corresponds to a Poisson error of about $\sqrt{10^{-8}} = 10^{-4}$ per bin).
2. There is ambiguity on how negative parallaxes (comprising 26% of the published *Gaia* DR2 data), which are perfectly valid measurements, should be treated. Again, depending on subjective choices, the observed distribution over the first significant digits changes by up to several percent points (see Appendix B and Fig. B.1).
3. In the same spirit, arbitrary changes of units (e.g. from parsec to light years [1 pc = 3.26 ly] or milliarcseconds to nanoradians [1 mas = 4.85 nrad]) change the observed distribution over the first significant digits by up to several percent points.
4. It is known that the *Gaia* DR2 parallaxes collectively have a global parallax zero-point offset, which to second order depends on magnitude, colour, and sky position (see Sect. 5 for a detailed discussion). Again, the absolute truth is out there, and depending on the subjective choice for the value of the offset correction, the observed distribution over the first significant digits changes significantly (see Fig. 14).

In short, even with a proper statistical test or metric, we would not be able to capture the effects of the existing freedom (and systematic effects) in the (interpretation of the) data.

Appendix E: ADQL queries

The following ADQL queries, and slight variations thereof, were used in this research:

– To query *Gaia* DR2 data from the *Gaia* Archive at ESA[4]:

```
SELECT BailerJones.source_id, BailerJones.r_est,
BailerJones.r_len, Gaia.parallax, Gaia.parallax_
over_error FROM external.gaiadr2_geometric_dista
nce as BailerJones INNER JOIN (SELECT GaiaData.
source_id, GaiaData.parallax, GaiaRUWE.ruwe FROM
gaiadr2.gaia_source as GaiaData INNER JOIN (SELECT
* FROM gaiadr2.ruwe where ruwe < 1.4) as GaiaRUWE
ON GaiaData.source_id = GaiaRUWE.source_id where
(GaiaData. phot_bp_rp_excess_factor < 1.3 + 0.06 *
POWER(GaiaData.phot_bp_mean_mag - GaiaData.phot_
rp_mean_mag, 2) and GaiaData.phot_bp_rp_excess_
factor > 1 + 0.015 * POWER(GaiaData.phot_bp_mean_
mag - GaiaData.phot_rp_mean_mag, 2))) as Gaia ON
BailerJones.source_id = Gaia.source_id;
```

– To query median StarHorse distance estimates for a random subset of high-quality objects from the *Gaia* Archive at the Leibniz Institute for Astrophysics in Potsdam (see Sect. 4.2)[5]:

```
SELECT TOP 1000000 StarHorse.source_id, StarHorse.
dist50, Gaia.parallax FROM gdr2_contrib.starhorse
as StarHorse INNER JOIN (SELECT source_id, paral
lax FROM gdr2.gaia_source ORDER BY random_index)
as Gaia ON StarHorse.source_id = Gaia.source_id
AND StarHorse.SH_OUTFLAG LIKE ''00000'' AND Star
Horse.SH_GAIAFLAG LIKE ''000''
```

– To query true GUMS distances from the *Gaia* Archive at the Centre de Données astronomiques de Strasbourg (Ochsenbein et al. 2000; see Sect. 4.4)[6]:

```
SELECT ''VI/137/gum_mw''.r FROM ''VI/137/gum_mw'',
```

after which a random sample of one million objects was selected using the shuf command.

– To query one million random observed GUMS parallaxes from the *Gaia* Archive at the Observatoire de Paris-Meudon (see Sect. 4.4)[7]:

```
SELECT parallax FROM simus.complete_source,
```

after which a random sample of one million objects was selected using the shuf command.

Appendix F: Selected mathematical derivations

For convenience of the reader, without pretending to have derived these relations as new discoveries, this appendix presents selected derivations linked to scale invariance (Appendix F.1), base invariance (Appendix F.2), and inverse invariance (Appendix F.3; see e.g. Hill 1995a and Weisstein 2019). Appendix F.4 discusses the distribution of first significant digits of a Lorentzian distribution.

[4] https://gea.esac.esa.int/archive/
[5] https://gaia.aip.de/query/
[6] http://tapvizier.u-strasbg.fr/adql/?gaia
[7] https://gaia.obspm.fr/tap-server/tap

F.1. Scale invariance

It is possible to define the probability for the first significant digit with a probability density function as follows:

$$P(D_1X = d) = P(\lfloor x \rfloor = d) = P(d \leq x < d+1) = \int_d^{d+1} p(x)\mathrm{d}x. \tag{F.1}$$

If Benford's law is a universal law, it needs to be independent of the selected unit (e.g. parsec or light year). In other words, the first significant digit distribution has to be scale invariant. If the distribution of data set $\tilde{X}$ is scale invariant, then there exists $\forall X \in \tilde{X}$, a scale $C \in \mathbb{R}_{>0}$, an $\alpha \in (0, 10)$, and a function f such that for a significand x of X, we have

$$P(D_1(CX) = d) = P(\lfloor CX \rfloor = d) = P(\lfloor \alpha x \rfloor = d) = P(\lfloor x \rfloor = d)/f(\alpha), \tag{F.2}$$

where

$$\alpha = \frac{c}{10} \quad \text{if the significand of } CX \text{ is smaller than the significand of } X; \tag{F.3}$$

$$\alpha = c \quad \text{if the significand of is larger than or equal to the } CX \text{ significand of } X, \tag{F.4}$$

such that

$$p(x) = f(\alpha) \cdot p(\alpha x). \tag{F.5}$$

We exclude $\alpha = C = 0$ because this is a special case for which Eq. (F.2) gives $d = 0$, when $P(D_1(0 \cdot X) = d)$ for every X.

In order to prove that Benford's law appears if and only if the data are scale invariant, we start with assuming that the data follow Benford's law. This means that the probability for the first significant digit d of significand x to appear equals

$$P(d \leq x < d+1) = \log_{10}\left(\frac{d+1}{d}\right) = \frac{1}{\ln(10)} \int_d^{d+1} \frac{1}{x}\mathrm{d}x. \tag{F.6}$$

Therefore the probability density function equals

$$p(x) = \frac{1}{\ln(10) \cdot x}. \tag{F.7}$$

This probability density function satisfies Eq. (F.5) for $f(\alpha) = \alpha$. This proves that if a data set follows Benford's law, it is scale invariant.

Next, we assume that the data are scale invariant. This implies that Eq.(F.5) holds for every $\alpha \in \mathbb{R}_{\geq 0}$. If $p(x)$ is a continuous probability density function on $[1, 10)$, such that

$$\int_1^{10} p(x)\mathrm{d}x = 1, \tag{F.8}$$

we can derive

$$P(1 \leq x < 10) = \int_1^{10} p(\alpha x)\mathrm{d}x = \int_1^{10} p(z)\frac{\mathrm{d}z}{\alpha} = \frac{1}{\alpha}, \tag{F.9}$$

where $z \equiv \alpha x$. This result gives $f(\alpha) = \alpha$. Therefore

$$\frac{p(x)}{\alpha} = p(\alpha x). \tag{F.10}$$

By taking the derivative of both sides with respect to α, and by choosing $\alpha = 1$, the following relation holds:

$$-p(x) = x\frac{\partial p(x)}{\partial x}. \tag{F.11}$$

This differential equation can be solved with the separation technique and gives

$$-\ln(x) + c = \ln(p(x)); \quad \frac{\lambda}{x} = p(x), \tag{F.12}$$

where $\lambda = e^c$. Now, it is possible to derive

$$\int_1^{10} p(x)\mathrm{d}x = \int_1^{10} \frac{\lambda}{x}\mathrm{d}x = \lambda\left[\ln(10) - \ln(1)\right] = 1, \tag{F.13}$$

so that

$$\lambda = \frac{1}{\ln(10)}, \tag{F.14}$$

and the probability density function $p(x)$ becomes

$$p(x) = \frac{1}{\ln(10) \cdot x}. \tag{F.15}$$

The probability of the first significant digit can now be derived by

$$P(D_1X = d) = P(d \leq x < d+1) = \int_d^{d+1} p(x)\mathrm{d}x = \frac{1}{\ln(10)} \int_d^{d+1} \frac{1}{x}\mathrm{d}x = \log_{10}\left(\frac{d+1}{d}\right). \tag{F.16}$$

This is exactly Benford's law for the first significant digit (Eq. (2)). This proves that if the data are scale invariant, they follow Benford's law. In conclusion, a necessary precondition for a data set to follow Benford's law is scale invariance.

F.2. Base invariance

If Benford's law is a universal law, it should be base invariant as well, next to being scale invariant, as these properties have a common origin[8]. Consider for example the scale invariance of a uniform logarithmic distribution, which shows that base invariance is related to scale invariance. We can generalise the scale-invariance derivation in Appendix F.1 by substituting base 10 with base B such that

$$\int_1^{B} p(x)\mathrm{d}x = \int_1^{B} \frac{\lambda}{x}\mathrm{d}x = \lambda\left[\ln(B) - \ln(1)\right] = 1. \tag{F.17}$$

Therefore we find

$$\lambda = \frac{1}{\ln(B)}, \tag{F.18}$$

which gives

$$P(D_1X = d) = \int_d^{d+1} p(x)\mathrm{d}x = \log_B\left(\frac{d+1}{d}\right). \tag{F.19}$$

Base invariance was discussed in detail by Hill (1995a).

[8] As already mentioned in Sect. 2, Hill (1995b) demonstrated that scale invariance implies base invariance (but base invariance does not imply scale invariance).

F.3. Inverse distribution

When parallaxes and distances are discussed, a relevant question is the relation between a data set $\tilde{X}$ and its inverse $\tilde{X}^{-1}$. From the scale invariance of a uniform logarithmic distribution, we might intuitively already expect that the inverse distribution is scale invariant as well. We here formally demonstrate that if $\tilde{X}$ is a data set that follows Benford's law, then the inverted data $\tilde{X}^{-1}$ also follow Benford's law.

First, we note that the mapping of the mantissae to the inverse of the mantissae is given by

$$\left\{\tilde{X} \to \tilde{X}^{-1} : x \mapsto x^{-1} \cdot 10^{b}\right\}, \tag{F.20}$$

where $b = 0$ if $x = 10^{n}\ \forall n \in \mathbb{Z}$ and $b = 1$ for any other value of x.

Next, we assume that $\tilde{X}$ follows Benford's law, such that for $\forall X \in \tilde{X}$ in significand notation $X = x \cdot 10^{m}$, with $m \in \mathbb{Z}$, we have

$$\begin{aligned} P(D_1X^{-1} = d) &= P(x^{-1} \in [d, d+1)) \\ &= P(d \leq x^{-1} < d+1) \\ &= P\left(\frac{10^{b}}{d+1} < x \leq \frac{10^{b}}{d}\right) \\ &= \frac{1}{\ln(10)} \int_{\frac{10^{b}}{d+1}}^{\frac{10^{b}}{d}} \frac{1}{x} \mathrm{d}x \\ &= \frac{1}{\ln(10)} \int_{\frac{1}{d+1}}^{\frac{1}{d}} \frac{1}{x} \mathrm{d}x \\ &= \frac{1}{\ln(10)} \left[-\ln(d) + \ln(d+1)\right] \\ &= \log_{10}\left(\frac{d+1}{d}\right). \end{aligned} \tag{F.21}$$

This result, together with scale invariance (Appendix F.1), implies that the mapping $X \mapsto \alpha X^{-1}$ preserves Benford's law for any value $\alpha \in \mathbb{R}_{>0}$.

F.4. Distribution of first significant digits of a Lorentzian distribution

The normalised Lorentzian function, centred at $x = x_0$ and with a half width at half maximum of γ, is given by

$$L(x; x_0, \gamma) = \frac{1}{\pi \cdot \gamma \left(1 + \left[\frac{x - x_0}{\gamma}\right]^2\right)}. \tag{F.22}$$

The first significant digit probability distribution for a Lorentzian function can be derived analytically by using the generic first significant digit probability function:

$$P(D_1X = d) = \sum_{k=-\infty}^{\infty} \int_{d \cdot 10^{k}}^{(d+1) \cdot 10^{k}} p(x) \mathrm{d}x, \tag{F.23}$$

and replacing the generic probability density function $p(x)$ with the Lorentzian from Eq. (F.22):

$$\begin{aligned} P(D_1X = d; x_0, \gamma) &= \frac{\sum_{k=-\infty}^{\infty} \int_{d \cdot 10^{k}}^{(d+1) \cdot 10^{k}} L(x; x_0, \gamma) \mathrm{d}x}{\int_{0}^{\infty} L(x; x_0, \gamma) \mathrm{d}x} \\ &= \sum_{k=-\infty}^{\infty} \left[\frac{\operatorname{atan}\left(\frac{(d+1) \cdot 10^{k} - x_0}{\gamma}\right) - \operatorname{atan}\left(\frac{d \cdot 10^{k} - x_0}{\gamma}\right)}{\frac{\pi}{2} + \operatorname{atan}\left(\frac{x_0}{\gamma}\right)}\right]. \end{aligned} \tag{F.24}$$

The normalisation assumes that only positive first significant digit values are considered ($x > 0$).

With the atan addition formula, given by

$$\begin{aligned} &\operatorname{atan}(x) + \operatorname{atan}(y) \\ &= \begin{cases} \operatorname{atan}\left(\frac{x+y}{1-xy}\right) & \text{if } xy < 1, \\ \operatorname{atan}\left(\frac{x+y}{1-xy}\right) + \pi & \text{if } xy > 1 \text{ and } x > 0 \text{ and } y > 0, \\ \operatorname{atan}\left(\frac{x+y}{1-xy}\right) - \pi & \text{if } xy > 1 \text{ and } x < 0 \text{ and } y < 0, \end{cases} \end{aligned} \tag{F.25}$$

and with the following two identities (valid for $x > 0$):

$$\operatorname{atan}(-x) = -\operatorname{atan}(x), \tag{F.26}$$

$$\operatorname{atan}\left(\frac{1}{x}\right) = \operatorname{acot}(x), \tag{F.27}$$

the first significant digit probability function can be written as

$$\begin{aligned} &P(D_1X = d; x_0, \gamma) \\ &= \sum_{k=-\infty}^{\infty} \left[\frac{\operatorname{acot}\left(\gamma \cdot 10^{-k} + x_0^2 \cdot 10^{-k} - (2d+1)x_0 + d(d+1)10^{k}\right)}{\frac{\pi}{2} + \operatorname{atan}\left(\frac{x_0}{\gamma}\right)}\right]. \end{aligned} \tag{F.28}$$

In view of Eq. (F.25), a minor modification of the above result is required for index $k = k^{*}$, defined as

$$d \cdot 10^{k^{*}} < x_0 < (d+1) \cdot 10^{k^{*}}, \tag{F.29}$$

and

$$\left((d+1) \cdot 10^{k^{*}} - x_0\right)\left(d \cdot 10^{k^{*}} - x_0\right) + \gamma^2 < 0, \tag{F.30}$$

such that the final result reads

$$\begin{aligned} &P(D_1X = d; x_0, \gamma) \\ &= \sum_{k=-\infty, k \neq k^{*}}^{\infty} \left[\frac{\operatorname{acot}\left(\gamma \cdot 10^{-k} + x_0^2 \cdot 10^{-k} - (2d+1)x_0 + d(d+1)10^{k}\right)}{\frac{\pi}{2} + \operatorname{atan}\left(\frac{x_0}{\gamma}\right)}\right] \\ &+ \sum_{k=k^{*}} \left[\frac{\operatorname{acot}\left(\gamma \cdot 10^{-k} + x_0^2 \cdot 10^{-k} - (2d+1)x_0 + d(d+1)10^{k}\right)}{\frac{\pi}{2} + \operatorname{atan}\left(\frac{x_0}{\gamma}\right)} + \pi\right]. \end{aligned} \tag{F.31}$$