Benford’s Distribution in Extrasolar World: Do the Exoplanets Follow Benford’s Distribution?

ABHISHEK SHUKLA1,*, ANKIT KUMAR PANDEY2 and ANIRBAN PATHAK1

1Jaypee Institute of Information Technology, A-10, Sector 62, Noida, 201 307, India.
2Indian Institute of Science Education and Research, Mohali 140 306, India.
*Corresponding author. E-mail: abhishek.shukla@mail.jiit.ac.in

MS received 9 August 2016; accepted 11 November 2016; ePublication: 15 March 2017

Abstract. In many real life situations, it is observed that the first digits (i.e., 1, 2, . . . , 9) of a numerical data-set, which is expressed using decimal system, do not follow a uniform distribution. In fact, the probability of occurrence of these digits decreases in an almost exponential fashion starting from 30.1% for 1 to 4.6% for 9. Specifically, smaller numbers are favoured by nature in accordance with a logarithmic distribution law, which is referred to as Benford’s law. The existence and applicability of this empirical law have been extensively studied by physicists, accountants, computer scientists, mathematicians, statisticians, etc., and it has been observed that a large number of data-sets related to diverse problems follow this distribution. However, except two recent works related to astronomy, applicability of Benford’s law has not been tested for extrasolar objects. Motivated by this fact, this paper investigates the existence of Benford’s distribution in the extrasolar world using Kepler data for exoplanets. The quantitative investigations have revealed the presence of Benford’s distribution in various physical properties of these exoplanets. Further, some specific comments have been made on the possible generalizations of the obtained results, its potential applications in analysing the data-set of candidate exoplanets.

Keywords. Benford’s distribution—exoplanets—significant digit law.

1. Introduction

In 1881, while going through the logarithms of an unbiased data-set, Simon Newcomb noticed an anomalous behavior in the distribution of digits (Newcomb 1881). Actually, he computed the occupancy of the most significant digit (MSD) from such a data-set. Counter to common intuition, which would expect an unbiased or uniform behavior in occupancy of the digits, Newcomb found that it decreases exponentially with digits. The probability of occurrence of 1 was found to be 30.1% and the same for 9 was found to be 4.6%. Newcomb’s prediction was empirical in nature, but due to lack of mathematical structure his article did not receive much attention. Later in 1938, Benford mathematically formulated a law to calculate probability $P_d$ of occurrence of the digit $d$ as the MSD, with the sum of the probability to be unity (i.e., $\sum_{d=1}^{9} P_d = 1$) (Benford 1938). The probability distribution introduced by Benford was

$$P_d = \log_{10} \left(1 + \frac{1}{d}\right).$$

Since the pioneering works of Newcomb and Benford, a large number of works related to Benford’s law have been reported in various contexts (for a fascinating history of Benford’s law, see Berger & Hill (2015, 2011), Adhikari & Sarkar 1968 and Berger et al. (2009). For example, its presence and applicability have been investigated in various domains, like astrophysics (Moret et al. 2006; Alexopoulos & Leontsinis 2014a, b; Hill & Fox 2016), geography (Sambridge et al. 2010), biology (Busta & Weinberg 1998; Cáceres et al. 2008; Docampo et al. 2009), seismography (Sottili et al. 2012), stock market and accounting (De Ceuster et al. 1998; Durtschi et al. 2004). Interestingly, violation of Benford’s law has been found to be capable of detecting cases of tax fraud (Busta & Sundheim 1992) and election fraud (Deckert et al. 2010), and it is routinely used by accounting professionals to detect financial irregularities. However, its reliability as a tool for fraud detection is still debatable. The issues and the cases where violation of Benford’s law do not correctly predict the presence of fraud are discussed in Busta & Sundheim (1992). It is not our purpose to discuss this
particularly interesting issue in detail, rather we are interested to note that Benford’s law-based analysis has recently drawn considerable attention of the physics community. Especially, after its formulation as an efficient tool to study quantum phase transitions (De & Sen 2011; Rane et al. 2014). Further, it has also been shown that Benford’s law-based analysis is helpful in spectroscopy. In particular, its applications for weak peak detection, phase correction and baseline correction have been demonstrated on NMR signal by some of the present authors (Bhole et al. 2015). A good agreement with ideal Benford’s distribution was observed in various NMR spectra, and that validated the existence of Benford’s distribution in NMR-based systems. Furthermore, an attempt to use Benford’s law-based analysis for processing MRI data has already been made in Bhole et al. (2015). This provides us an excellent example of the application of Benford’s law. In addition, in an attempt to reveal the existence of first-principle-based rule behind Benford’s law, Shao & Ma (2010) have reported Benford’s distribution for various statistical ensembles. They found that Maxwell–Boltzmann and Bose–Einstein statistics allow periodic fluctuations in occupancy of digits with temperature, while for Fermi–Dirac statistics such fluctuations remain absent (Shao & Ma 2010).

Until the recent past, all the investigations on Benford’s law were restricted to the data-set generated in the context of our solar system, in general, and the Earth, in particular. However, recent astrophysical observations reported in Moret et al. (2006), Alexopoulos & Leontsinis (2014a, b) and Hill & Fox (2016) have established that Benford’s law is followed by star distances and distances from the Earth to galaxies. This observation, and the fact that one of the most prominent interest of mankind is to find promising sites to host an extra-terrestrial life, have motivated us to ask: “Do exoplanets follow Benford’s distribution?” We try to answer this particular question using Kepler data (kep 2016), which provides various information related to exoplanets that are mainly detected by NASA’s Kepler telescope. The size of this data-set (i.e., Kepler data) has been considerably increased recently as NASA has confirmed the existence of several exoplanets. With this new announcement, the number of detected and confirmed exoplanets goes to \( \approx 3300 \) which is of reasonable size for statistical analysis of the data-set, in general, and for investigation on the existence of Benford’s distribution, in particular. This point would be more clear if we note that in Pintr et al. (2013, 2014), the statistical analysis of exoplanets data was performed by some of the present authors using a data-set of (1771) exoplanets, which was the number of exoplanets known at that time. Still, Pintr et al. (2014) yielded various interesting results related to the possibility of existence of habitable exoplanets (nat 2014).

The remaining part of this paper is organized as follows. In section 2, we briefly describe the method adopted here for the investigation of Benford’s distribution, and the quantitative measures used for comparing the similarity between the Benford’s distribution, and the actual distribution. In section 3, we describe our results. Finally, we conclude section 4, where we have also mentioned some potential applications of the present work.

2. Method

To calculate Benford’s distribution for a given data-set we have adopted the simplest method described in (Simkin 2010), where it is shown that the distribution of MSD can be obtained using a spread-sheet (Microsoft Excel or a similar program). Using the above mentioned procedure, we have calculated probability of occurrence for each digit in the Kepler data for exoplanets (kep 2016). The same is illustrated through a set of plots. Specifically, in Fig. 1, we illustrate the distribution of MSDs for quantities mass, volume, density, orbital semi major axis, orbital period and radial velocity in Kepler data for exoplanets. All sub-plots of Fig. 1 clearly show that the values of a set of physical properties (e.g., mass, volume, density, orbital semi-major axis, orbital period, and radial velocity) are distributed in a manner that nicely matches with Benford’s distribution. In other words, Fig. 1 shows that exoplanets follow Benford’s distribution. However, in all the subplots, the matching between the real distribution and the ideal Benford’s distribution is not the same. Thus, to understand how closely the values associated with a particular property follow Benford’s distribution, we need a quantitative measure. Such a quantitative measure is \( \chi^2 \), which implicate solid statistical conclusions. For a given data-set, \( \chi^2 \) is defined as

\[
\chi^2 = \sum_{d=1}^{9} \left(\frac{P(d) - P_B(d)}{P_B(d)}\right)^2.
\]

Here, \( P(d) \) is the observed probability for digit \( d \) and \( P_B(d) \) is the ideal probability in Benford’s distribution for the same digit \( d \). A complete overlap corresponds to \( \chi^2 = 0 \), the greater value of \( \chi^2 \) for a data-set implies relatively bad overlap of actual distribution with
Benford’s distribution. In other words, we may use \( \chi^2 \) as a quantitative measure of how accurately a given data-set follows Benford’s distribution. In practice, \( \chi^2 \) analysis is actually used to check whether a dataset follows a given distribution by making a null hypothesis and an alternative hypothesis, and subsequently computing the \( p \) value. Specifically, for our case, the null hypothesis \( H_0 \) would be stated as the first digits of the property \( S \) of exoplanets that follow Benford’s distribution, and the alternative hypothesis \( H_1 \) would be stated as the first digits of the property \( S \) of exoplanets that do not follow Benford’s distribution, where \( S \in \{ \text{mass, volume, density, orbital semi-major axis, orbital period and radial velocity} \} \) of exoplanets. A subsequent computation of \( p \) value would establish if we obtain \( p > \alpha = 0.05 \) (\( p \leq \alpha = 0.05 \)), then the validity of \( H_0(H_1) \) cannot be rejected (value of significance level \( \alpha \) is usually selected as 0.05. This value of chosen significance level is standard, but one may decide to choose another value). For the present study \( \alpha \) is chosen as 0.05, degrees of freedom is 8, which is fixed for all the data-sets (properties) studied here, and \( p \) value for a specific property of exoplanet is computed using \( \chi^2 \) value obtained for that property and degrees of freedom with the help of Microsoft Excel program. For a perfect match \( p \) value would be 1. Actually \( p \) approaches 1 for almost perfect matches, too. In what follows, we report \( \chi^2 \) for all the properties of exoplanets. To be precise, in the following section, we have used \( \chi^2 \) measure to analyse Kepler data for exoplanets (kep 2016). Specifically, data for density, orbital period and orbital semi-major axis were taken from (kep 2016) on May 17, 2016, while data for the rest of the quantities have been taken from the same data archive on May 04, 2016.

Driven by the curiosity of examining deeper statistical symmetry present in Kepler data for exoplanets, we have also calculated joint probability of occupancy \( P(d_1, d_2) \) for the first and second significant digits being \( d_1 \) and \( d_2 \), respectively. For the purpose, we have used Hill’s formula (Hill 1995) for generalized Benford’s distribution, which states that the probability \( P(d_1, d_2, \ldots d_N) \) for digits \( d_1, d_2, \ldots d_N \) is

\[
P(d_1, d_2, \ldots d_N) = \log_{10} \left[ 1 + \left( \sum_{i=1}^{k} d_k 10^{k-i} \right)^{-1} \right].
\]  

In particular, using equation (3), we have calculated \( P(d_1, d_2) \) for some quantities namely mass, volume, orbital period, effective temperature and radius. We found encouraging results in case of orbital period, but not in other cases. In the next section, we will discuss these results in detail.

### 3. Result and Discussion

Figure 1 illustrates plots for observed and ideal Benford’s distribution for various physical quantities, namely mass, density, orbital period, volume and orbital semi-major axis of exoplanets. The \( \chi^2 \) and \( p \) values for mass (0.027, 0.999999999), density (0.014, 0.999999999907), orbital period (0.011, 0.999999999959), volume (0.025, 0.99999998849), and orbital semi-major axis (0.024, 0.999999999021) are given in the bracket against a property respectively. As we obtain very high values of \( p \) (i.e., \( p >> 0.05 \)) for the properties discussed above, we cannot reject corresponding null hypothesis and conclude that first digits of these properties significantly follow Benford’s distribution. From Fig. 1, we find that orbital period of exoplanets most closely follows Benford’s distribution as \( \chi^2 \) for this set of data is 0.011 (cf. Fig. 1(a)). Here, it may be noted that in Fig. 1(d), the radius used for calculating the volume is absolute, as it has been obtained by multiplying the radius of the Earth into the data obtained from the Kepler archive. Similarly, in Fig. 2(d), the radius for calculating the volume is absolute, as it has been obtained by multiplying the radius of the Earth into the data obtained from the Kepler archive. However, these scalings have not affected the distribution of MSDs, as Benford’s distribution is known to be scale-independent (Miller 2015). Further, from Fig. 1, we observe that the exoplanet’s orbital period (cf. Fig. 1(a)), density (cf. Fig. 1(b)), orbital semi-major axis (cf. Fig. 1(c)), volume (cf. Fig. 1(d)), and mass (cf. Fig. 1(e)) nicely follow Benford’s distribution. Specifically, \( \chi^2 \), computed for all these physical properties are less than 0.04. However, the Kepler data for other physical properties do not follow Benford’s law so strictly. To be precise, we have observed that there are some quantities, which have only moderate overlap with the ideal Benford’s distribution. These quantities are stellar distance, total proper motion, radial velocity and stellar age of exoplanets. In fact, the quantitative measure using \( \chi^2 \) allows us to construct a criteria for classifying data. Specifically, we consider that \( \chi^2 \) values \( \leq 0.04 \) correspond to a good agreement, \( 0.04 \leq \chi^2 \leq 0.13 \) corresponds to intermediate agreement, and \( \chi^2 \geq 0.13 \) implies bad agreement.

Now, we may look at Fig. 2, which contains observed and ideal Benford’s distribution plots for various physical quantities that are not illustrated in the previous
Figure 1. (Color online). Figure contains probability (vertical axis) of occurrence of MSD (horizontal axis). An ideal Benford’s distribution is shown in orange color and the observed probability distributions are shown in blue color. Subplots (a)–(e) show Benford’s distributions of MSD for various physical quantities obtained from the Kepler archive (kep 2016). In each subplot corresponding $\chi^2$ value is noted.

Figure 2. (Color online). Figure contains probability (vertical axis) of occurrence vs. MSD (horizontal axis). Similar to the previous figure, ideal Benford’s distribution is shown in orange color and the observed probability distribution is shown in blue color. The top and middle trace contain those physical quantities who have $0.04 \leq \chi^2 \leq 0.13$ whereas, the bottom trace incorporates those physical quantities which have $\chi^2 \geq 0.13$. The physical quantity and the corresponding $\chi^2$ value of the distribution are mentioned in every subplot.
Figure 3. (Color online). Figure contains probability (vertical axis) vs. digits $d_1d_2$ (horizontal axis). Again ideal Benford’s distribution is shown in orange color and observed probability distribution is shown in blue color. Upper panel shows results on a data-set of 1898 exoplanets while lower panel illustrates the results for an enlarged data-set of 3207 exoplanets. The corresponding $\chi^2$ values are also mentioned (0.06 for 1898 exoplanets and 0.036 for 3207 exoplanets).

plot. Figures 2(a)–(d) elucidate distribution of first digits for all those quantities that moderately follow Benford’s distribution. The corresponding $\chi^2$ and p values are (0.046, 0.99999988230722), (0.05, 0.99999984610641), (0.064, 0.99999956000) and (0.093, 0.99999973050787). Similarly, Figures 2(e)–(g) illustrate the statistical distribution of the MSD for longitude (in radians), radius and effective temperature (wik 2015), and it is observed that data-set for these properties do not follow Benford’s distribution. The same is also quantitatively reflected in the $\chi^2$ values ($\chi^2 \geq 0.130$) obtained for longitude, radius and effective temperature data-set. On keen observation, we noticed that those quantities which have $\chi^2 \geq 0.130$ were actually having small variation in data, typically they are of the same order (or variation in orders of magnitude is very narrow). Such a data-set may be considered as biased data-set. This explains why Benford-like distribution is not observed in Figures 2(e)–(g) (i.e., for longitude, radius and effective temperature of exoplanets). Thus, in brief, we may state that except a few incidents of biased data, it is observed, in general, that Benford’s distribution is followed by the values of most physical properties associated with the exoplanets.

Inspired by the observation that MSDs of values of various properties associated with exoplanets follow Benford’s distribution, we tried to investigate whether the second MSDs also follow this distribution. To do so, we have computed $P(d_1, d_2)$ using equation 3 for a few physical properties (e.g., orbital period, mass and volume). Here, we illustrate our observations only on the orbital period. In Fig. 3, the overlap of $P(d_1, d_2)$, obtained from the real data and from the ideal Benford’s distribution is shown. It is observed that $\chi^2$ increases with the size of data-set. In particular, for a data-set of 1898 exoplanets, we obtained $\chi^2 = 0.06$, and for a data-set of 3207 exoplanets we obtained $\chi^2 = 0.036$. Surprisingly, we did not observe this increase in $\chi^2$ with size of the data-set for other physical properties (mass and volume). A probable reason for this observed increase in $\chi^2$ of two digit distribution $P(d_1, d_2)$ for orbital period, but not in the case of other properties may be that the data-size of the second data-set is sufficiently large to yield unbiased nature of data so that it can follow two-digit Benford’s distribution and hence become close to ideal Benford’s distribution. In contrast, for other properties, may be the data-size is still not sufficient to realize an unbiased two-digit data-set.
4. Conclusions

The validity of Benford’s law is investigated for the first time for exoplanets. The investigation is performed using Kepler data, and it is observed that the data-set corresponding to exoplanet’s orbital period, density, orbital-semi-major axis, mass and volume nicely follow this law, whereas exoplanet’s stellar distance, radial velocity and stellar age moderately follow Benford’s law. On the other hand, exoplanet’s longitude, radius, and effective temperature hardly follow the law. This is illustrated through Figures (1)–(3), and is clearly established by the quantitative measures $\chi^2$ and $p$. It is found that the $\chi^2$ is minimum for orbital period of the exoplanets (0.011), and is quite less for the density of the exoplanets (0.013) too. Thus, these two parameters almost exactly follow Benford’s distribution as the complete overlap corresponds to 0 $\chi^2$ value. Such a good overlap is rare for any data-set, and this observation has provided us a clear affirmative answer to the question that we asked in the beginning: Do the exoplanets follow Benford’s distribution?

Statistical distribution of exoplanet’s orbital period and density clearly established the fact that exoplanets follow Benford’s distribution, and validity of Benford’s law holds even for exoplanets. This strong observation is further supported by the fact that exoplanet’s orbital semi-major-axis ($\chi^2 = 0.024$), volume ($\chi^2 = 0.025$) and mass ($\chi^2 = 0.027$) also strongly follow Benford’s law. Thus, this empirical law seems to be universal and probably it is more fundamental and profound in nature than it is understood to be. However, some questions are still open. Why it works for certain data-set, and why it does not work for others? What is obtained until now is a mathematical insight that help us to understand where (i.e., in which data-sets) it works and where it does not. For example, we know that a data-set which is not biased and where the order of magnitude varies considerably, is expected to follow this law. This point only answer ‘where’ or ‘when’, but neither provides any physical insight (an understanding from the first principle) nor answer ‘why’. Thus, it needs more investigations. One may also generalize the results easily. For example, one may examine whether the Benford’s distribution is followed by second, third, fourth,... significant digits by using a general version of Benford’s distribution introduced by Hill (1995).

Recently, equation (3) has been used in Alexopoulos & Leontsinis (2014a) to establish that MSD for star distances agrees very well with the Benford’s distribution as far as the first, second and third significant digits are concerned. Similar exercises can be performed using Kepler data and other available data-sets of interest. Motivated by this fact, we have computed $P(d_1, d_2)$ using equation (3) for a set of physical properties. Another version of Benford’s law has been introduced in chapter 1 of Miller (2015), and the authors have referred to it as strong Benford’s law, which provides the probability of a specific number ‘$x$’ in a data-set (see Definition 1.4.3 in Miller (2015)). Thus, the presented result can be generalized, and similar results can be obtained in other data-sets of interest. However, before we perform such an exercise, we must ask whether it is worthy to perform such an exercise? Whether such an investigation is expected to provide some physical information or new insights to the data-set. The answer is yes, and in what follows, we elaborate this by discussing a specific possibility.

We have already mentioned that Benford’s distribution has been successfully used in accounting to detect frauds (Busta & Sundheim 1992), which may be viewed as a noise introduced by a person or a group of persons in a data-set, which was otherwise expected to follow Benford’s distribution. Now, the present paper establishes that several physical parameters associated with the exoplanets nicely follow Benford’s distribution. This implies that in analogy with accounting, we may try to locate noise (which is analogous to fraud in accounting) in the Kepler data-set of candidate exoplanets (a set of potential exoplanets whose status are not yet confirmed), and that can ease our effort to locate actual exoplanets.

Finally, we would like to note that statistical analysis of Kepler data is not new. Earlier studies performed by some of us (Pintr et al. 2014) had revealed the region, where to look for habitable exoplanets, and the present study hints for a method to analyze candidate exoplanets. Keeping all these in mind, we conclude the paper optimistically, with a hope that this work would lead to a few more statistical investigations in similar directions and those investigations would provide more physical insights on: Why does Benford’s law work universally?

Acknowledgements

AP and AS thank Defense Research & Development Organization (DRDO), India for the support provided through the project number ERIP/ER/1403163/M/01/1603. AKP thanks JIIT, Noida (where this work is done) for the hospitality and facilities provided during his visit as a summer intern.
References

Adhikari, A., Sarkar, B. 1968, *Sankhyā: Ind. J. Stat. Series B*, 47.
Alexopoulos, T., Leontsinis, S. 2014a, *J. Astrophys. Astr.*, 35, 639.
Alexopoulos, T., Leontsinis, S. 2014b, arXiv preprint arXiv:1401.5794.
Benford, F. 1938, *Proc. A Phil. Soc.*, 78, 551.
Berger, A., Hill, T. P., Rogers, E. 2009, Benford online bibliography, http://www.benfordonline.net/list/chronological.
Berger, A., Hill, T. P. 2011, *Probab. Surv.*, 8, 1.
Berger, A., Hill, T. P. 2015, *An introduction to Benford’s Law*, Princeton University Press.
Bhole, G., Shukla, A., Mahesh, T. S. 2015, *Chem. Phys. Lett.*, 639, 36.

Busta, B., Sundheim, R. 1992, Center for Business Research Working Paper, 94.
Busta, B., Weinberg, R. 1998, *Manag. Audit. J.*, 13, 356.
Cáceres, J. L. H., García, J. L. P., Martínez Ortiz, C. M., Domínguez, L. G. 2008, *Electron. J. Biomed.*, 1, 27.
De, A. S., Sen, U. 2011, *Europhys. Lett.*, 95, 50008.
De Ceuster, M. J., Dhaene, G., Schatteman, T. 1998, *J. Emp. Fin.*, 5, 263.
Deckert, J., Myagkov, M., Ordeshok, P. C. 2010, Caltech/MIT Voting Technology Project Working Paper.
Docampo, S., del Mar Trigo, M., Aira, M. J., Cabezudo, B., Flores-Moya, A. 2009, *AeroBiologia*, 25, 275.

Durtschi, C., Hillison, W., Pacini, C. 2004, *J. Foren. Account., 5*, 17.
Hill, T. P. 1995, *Proc. Amer. Math. Soc.*, 123, 887.
Hill, T. P., Fox, R. F. 2016, *J. Astrophys. Astr.*, 37, 1.
Kep 2016, Exoplanet data archive for confirm planets, http://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=planets.
Miller, S. J. (ed.) 2015, *Benford’s law: Theory and applications*, Princeton University Press.
Moret, M. A., de Senna, V., Pereira, M. G., Zebende, G. F. 2006, *Int. J. Modern Phys. C*, 17, 1597.
Nat 2014, Looking for habitable planets beyond solar system, http://www.natureasia.com/en/nindia/article/10.1038/nindia.2014.123.
Newcomb, S. 1881, *A. J. Math.*, 4, 39.
Pintr, P., Peřinová, V., Lukš, A., Pathak, A. 2013, *Planet. Space Sci.*, 75, 37.
Pintr, P., Peřinová, V., Lukš, A., Pathak, A. 2014, *Planet. Space Sci.*, 99, 1.
Rane, A. D., Mishra, U., Biswas, A., Sen, A., Sen, U. et al. 2014, *Phys. Rev. E*, 90, 022144.
Sambridge, M., Tkalčič, H., Jackson, A. 2010, *Geophys. Res. Lett.*, 37.
Shao, L., Ma, B.-Q. 2010, *Phys. A: Stat. Mech. Appl.*, 389, 3109.
Simkin, M. G. 2010, *ISACA J.*, 1, 47.
Sottili, G., Palladino, D. M., Giaccio, B., Messina, P. 2012, *Math. Geosci.*, 44, 619.
Wik 2015, Effective temperature, https://en.wikipedia.org/wiki/Effective_temperature.