Chirag Gokani

AP Statistics Project 3 Due May 1, 2017 100%

This project counts as a test grade.

Due April 18:

Turn in a hard copy of a published article on an appropriate topic of interest to you. 4%

This article should:

--have been published since May 1, 2016 4%

--contain statistical terms which you have highlighted 4%

--include references to an observational study or an experiment 4%

--clearly include its source and date of publication 4%

Due on or before May 1:

Turn in a folder containing both a hard copy of your highlighted article and a copy of your document written on a word processor.

Your document should:

--have your name on each page 2%

--state the name of the article, its author, its source, and the date of publication 2%

--list the statistical terms in your selected document, along with

a) a rigorous definition of each term and 12%

b) an interpretation of each term in the context of your article.

--be written with complete sentences 2%

--describe in context and detail the study or experiment discussed 10%

--explain why you chose this topic and/or article 10%

--describe topics/procedures we studied in this course that were identified in the report 10%

--comment on whether or not the report was prepared accurately, from both information gathering and reporting standpoints. 10%

--describe your reactions to the results described in the article. 10%

--explain any suggestions you have for improvement of the writing of the article or the statistical methods used to reach a conclusion 10%

--be written on only one side of the paper. 2%

WOW! Impressive! Well done!

Benford's distribution in extrasolar world: Do the exoplanets follow Benford's distribution?

Abhishek Shukla[1], Ankit Kumar Pandey[2], and Anirban Pathak[1]*
[1] *Jaypee Institute of Information Technology, A-10, Sector 62, Noida, UP 201307, India*
[2] *Indian Institute of Science Education and Research, Mohali 411008, India*

In many real life situations, it is observed that the first digits (i.e., $1, 2, \ldots, 9$) of a numerical data-set, which is expressed using decimal system, do not follow a random distribution. Instead, the probability of occurrence of these digits decreases in 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 investigation has revealed the presence of Benford's distribution in various physical properties of these exoplanets. Further, Benford goodness parameters are computed to provide a quantitative measure of coincidence of real data with the ideal values obtained from Benford's distribution. The quantitative analysis and the plots have revealed that several physical parameters associated with the exoplanets (e.g., mass, volume, density, orbital semi-major axis, orbital period, and radial velocity) nicely follow Benford's distribution, whereas some physical parameters (e.g., total proper motion, stellar age and stellar distance) moderately follow the distribution, and some others (e.g., longitude, radius, and effective temperature) do not follow Benford's distribution. Further, some specific comments have been made on the possible generalizations of the obtained result, its potential applications in analyzing data-set of candidate exoplanets, and how interested readers can perform similar investigations on other interesting data-sets.

Keywords: Benford's distribution, exoplanets, Benford goodness parameter

I. INTRODUCTION

In 1881, while going through the logarithms of an unbiased data-set, Simon Newcomb noticed an anomalous behavior in the distribution of digits [1]. Actually, he computed occupancy of most significant digit (MSD) from such a data-set. Counter to common intuition, which would expect an unbiased or random 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 % for 1 and the same for 9 was found to be 4.6 %. Simon's prediction was empirical in nature, and due to lack of mathematical structure his article did not receive much attention. Later in 1938, Benford, (see image shown in Fig. 1) 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., $\sum_1^9 P_d = 1$) [2]. The probability distribution introduced by Benford was

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

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 [3–5]). For example, it's presence and applicability have been investigated in various domains, like astrophysics [6, 7], geography [8], biology [9–11], seismography [12], stock market and accounting [13, 14]. Interestingly, violation of Benford's law has been found to be capable of detecting cases of tax fraud [15] and election fraud [16], and it's 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 presence of fraud are discussed in [15]. 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 physics

* anirban.pathak@gmail.com

FIG. 1. A pencil sketch of Frank Benford.

community. Especially, after it's formulation as an efficient tool to study quantum phase transitions by De and Sen [17] and Rane et al., [18]. 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 [19]. 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 law-based analysis for processing MRI data has already been made in Ref. [19]. This provides us an excellent example of application of Benford's law. In addition, in an attempt to reveal the existence of first-principle-based rule behind Benford's law, Shao et al., have reported Benford's distribution for various statistical ensembles [20]. 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 remains absent [20].

Until recent past, all the investigations on the Benford's law were restricted to the data-set generated in context of our solar system in general, and the Earth in particular. However, recent astrophysical observations reported in [6] and [7] 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 [21], which provides various information related to exoplanets that are mainly detected by NASA's Kepler telescope. 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 a 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 Ref. [22, 23], the statistical analysis of exoplanets data was performed by some of the present authors using a data-set of (1771) exoplants, which was the number of exoplanets known at that time. Still, Ref. [23] yielded various interesting results related to the possibility of existence of habitable exoplanets [24].

Remaining part of this paper is organized as follows. In Sec. II, we briefly describe the method adopted here for the investigation of Benford's distribution, and the method adopted for computing Benford goodness parameter (BGP), which may be viewed as a measure of similarity between the Benford's distribution, and the actual distribution. In Sec. III, we describe our results. Finally, we conclude the paper in Sec. IV, where we have also mentioned some potential applications of the present work.

II. METHOD

To calculate Benford's distribution for a given data-set we have adopted the simplest method described in Ref. [25], 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 [21]. The same is illustrated through a set of plots. Specifically, in Fig. 2 we illustrate the distribution of MSDs for Kepler data for exoplanets. All sub-plots of Fig. 2, 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. 2 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. Interestingly, such a quantitative measure exists [26], and referred to as the BGP. For a given data-set, BGP is defined as

$$\mathrm{BGP} = \Delta\mathrm{P} = 100\left(1 - \sqrt{\sum_{\mathrm{d}=1}^{9}\frac{(\mathrm{P(d)} - \mathrm{P_B(d)})^2}{\mathrm{P_B(d)}}}\right). \quad (2)$$

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 $\Delta P = 100$, but there is no lower limit. Thus, the larger the value of ΔP or BGP for a data-set, the closer it is to the ideal Benford's distribution. In other words, we may use ΔP or BGP as a quantitative measure of how accurately a given data-set follows Benford's distribution. In the following section we have used this measure to analyze Kepler data for exoplanets [21]. Specifically, data for density, orbital period, and orbital semi-major axis were taken from [21] 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 first and second significant digits being d_1 and d_2, respectively. For the purpose, we have used Hill's [27] formula for generalized Benford distribution, which states that the probability $P(d_1, d_2, \cdots d_N)$ for digits d_1, d_2, $\cdots$ d_N is

$$P(d_1, d_2, \cdots d_N) = \log_{10}\left[1 + \left(\sum_{i=1}^{k} d_k 10^{k-i}\right)^{-1}\right]. \quad (3)$$

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

III. RESULT AND DISCUSSION

Fig. 2 illustrates plots for observed and ideal Benford's distribution for various physical quantities, namely mass, density, orbital period, volume, orbital semi-major axis, and radial velocity of exoplanets with BGP values 99.92, 83.92, 89.40, 83.92, 84.24, 99.58, respectively. From Fig. 2, we find that mass of exoplanets most closely follows Benford's distribution as BGP for this set of data is 99.92 (cf. Fig. 2 a). Here, it may be noted that in Fig. 2 a, the probabilities of occurrence of the MSDs are computed after multiplying Jupiter mass into the data obtained from the Kepler archive. Similarly, in Fig. 3 d, radius used for calculating 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 effected the distribution of MSDs, as Benford's distribution is known to be scale-independent. Further, from Fig. 2 we observe that the exoplanets' mass (cf. Fig. 2 a), density (cf. Fig. 2 b), orbital period (cf. Fig. 2 c), volume (cf. Fig. 2 d), orbital semi-major axis (cf. Fig. 2 e), and radial velocity (cf. Fig. 2 f) nicely follow Benford's distribution. Specifically, BGP computed for all these physical properties are greater than 80. However, the Kepler data for other physical properties don't 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 total proper motion, stellar age and stellar distance of exoplanets. In fact, quantitative measure of BGP allows us to construct a criteria for classifying data depending on their BGP values. Specifically, we consider that BGP values greater than 80 correspond to a good agreement, BGP values in the range ($60 < \mathrm{BGP} \leq 80$) correspond to moderate or intermediate agreement, and BGP values ≤ 60 implies that data-set don't follow Benford's distribution or equivalently, bad agreement. Now, we may look at Fig. 3, which contains observed and ideal Benford's distribution plots for various physical quantities that are not illustrated in the previous plot. Fig. 3 (a-c) illustrate distribution of first digits for all those quantities that moderately follow Benford's distribution (thus, their BGP values are in the range $60 < \mathrm{BGP} \leq 80$). This set includes total proper motion, stellar age, and stellar distances, and corresponding BGP values are 77.74, 69.44, and 78.47, respectively. Similarly, Fig. 3 (d-f) illustrate the statistical distribution of most significant digits for values of longitude (in radians), radius, and effective temperature [28], and it is observed that data-set for these properties do not follow Benford's distribution. The same is also quantitatively reflected in the BGP values ($\mathrm{BGP} \leq 60$) obtained for longitude, radius and effective temperature. On keen observation, we noticed that those quantities who do not have good BGP values were actually having small variation in data, typically they are of the same order (or variation in orders of magnitude is very narrow) and hence have lower BGP. Such a data-set may be considered as biased data-set. This explains, why Benford-like distribution is not observed in Fig. 3 (d-f) (i.e., for

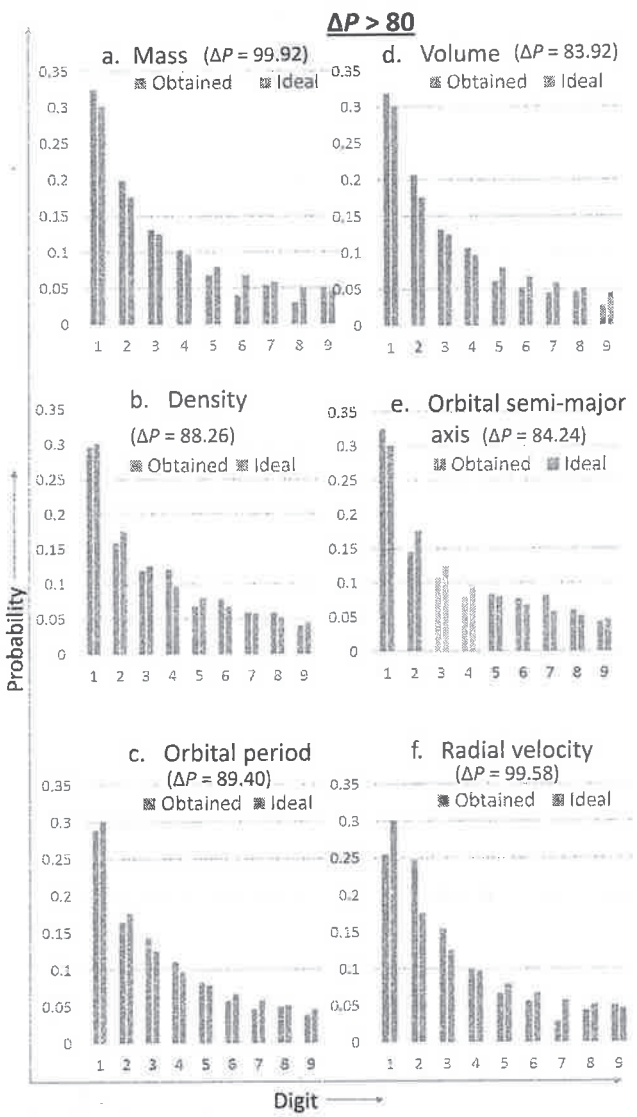


FIG. 2. (Color online) Figure contains probability (vertical axis) of occurrence of MSD (horizontal axis). Ideal Benford's distribution is shown in orange color and the observed probability distributions are shown in Blue color. Subplots (a-f) show Benford's distributions of MSD for various physical quantities obtained from the Kepler archive [21]. In each subplot corresponding BGP value is noted.

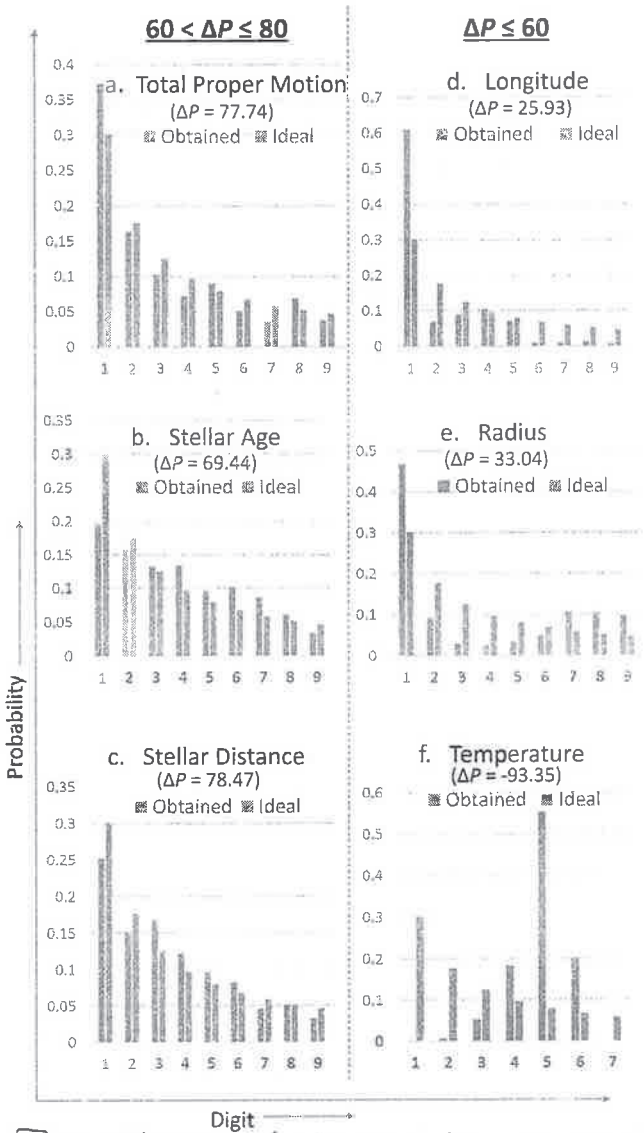


FIG. 3. (Color online) Figure contains probability (vertical axis) of occurrence vs MSD (horizontal axis). Similar to previous figure, ideal Benford's distribution is shown in orange color and the observed probability distribution is shown in Blue color. Left column contains those physical quantities who have BGP, $60 < \Delta P \leq 80$, whereas, right column incorporates those physical quantities who have BGP ≤ 60. The physical quantity and the corresponding BGP value of the distribution are noted in every subplot.

longitude, radius and effective temperature of exoplants). Thus, in brief, we may state that leaving 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 for values of various properties associated with exoplanets follow Benford's distribution, we tried to to investigate whether the second MSDs also follow this distribution. To do so, we have computed $P(d_1, d_2)$ using Eq. 3 for a few physical properties (e.g., orbital period, mass and volume). Here, we illustrate our observations on orbital period only. In Fig. 4, the overlap of $P(d_1, d_2)$ obtained from real data and ideal Benford's distribution is shown. It's observed that BGP increases with the size of data-set. In particular, for a data-set of 1898 exoplanets, we obtained

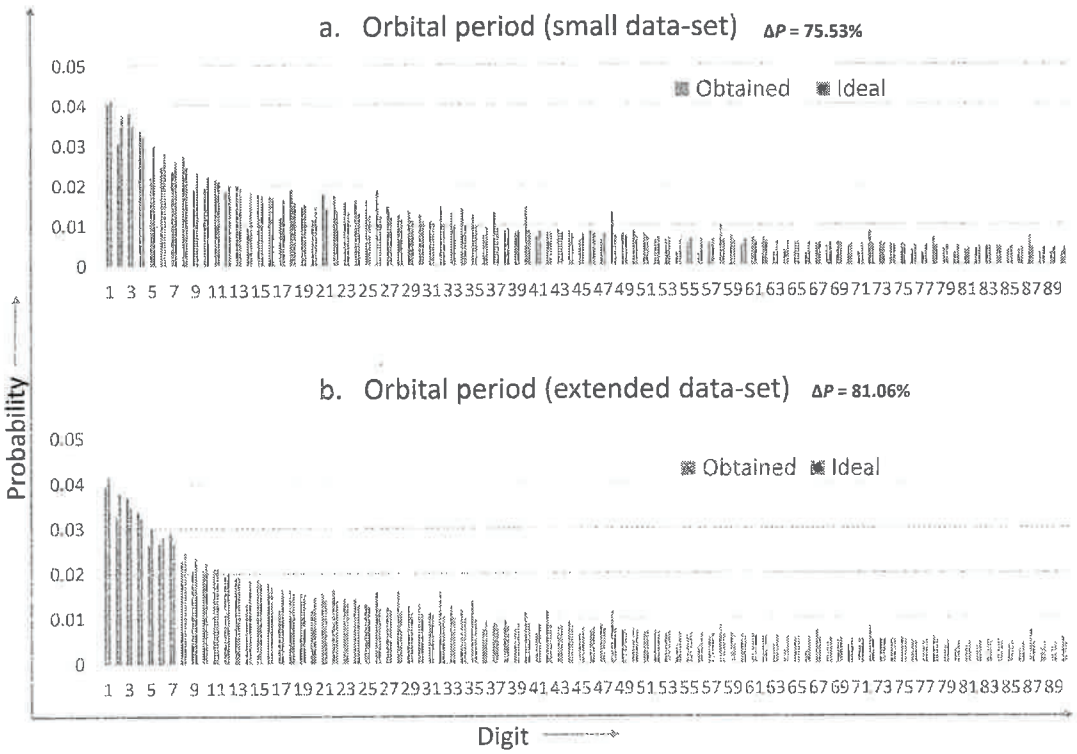


FIG. 4. (Color online) Figure contains probability (vertical axis) vs first and second significant digits d_1d_2 (horizontal axis). Again ideal Benford's distribution is shown in orange color and observed probability distribution is shown in blue color. Upper panel shows results on a data-set of 1898 exoplanets while lower panel illustrates the results for an enlarged data-set of 3207 exoplanets. Corresponding BGPs are also mentioned (75.53 for 1898 exoplanets and 81.06 for 3207 exoplanets).

BGP= 75.53, and for a data-set of 3207 exoplanets we obtained BGP=81.06. Surprisingly, we did not observe this increase in BGP with size of data-set for other physical properties (mass and volume). A probable reason for this observed increase in BGP of two digit distribution $P(d_1, d_2)$ for orbital period, but not in the case of other properties may be that for the former case data-size is sufficiently large to yield unbiased nature of data so that it can follow two digit Benford distribution and hence becomes close to ideal Benford distribution. In contrast, for other properties, may be the data-size is still not sufficient for realization of unbiased two digit data-set.

IV. 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 exolanets mass, volume, density, orbital semi-major axis, orbital period and radial velocity nicely follow this law, whereas exoplanets' total proper motion, stellar age, and stellar distance moderately follow Benford's law, and exoplanets' longitude, radius, and effective temperature hardly follow the law. This is illustrated through Figs. 2 - 4, and clearly established by a quantitative measure called BGP. It is found that the BGP is highest for the mass of the exoplanets (99.92), and it is extremely high for the radial velocity of the exoplanets (99.58), too. Thus, these two parameters almost exactly follow Benford's distribution as the upper-bound for BGP is 100. Such a high BGP 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 exoplanets' mass and radial velocity clearly established the fact that exoplanets follow Benford's distribution, and validity of Benford's law is not restricted to Earth, rather it's universal as it's followed in extrasolar worlds, too. This strong observation is further supported by the fact that exoplanets' density (BGP=88.26), orbital periods (BGP=89.40), volume (BGP=83.92), and orbital semi-major axis (BGP=84.24) also strongly follow Benford's law. Thus, this empirical

law seems to be universal and probably it's more fundamental and profound in nature than it's understood to be. However, some questions are still open. It's not well understood (physically), 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 helps us to understand where (i.e., in which data-sets) it works and where it does not. For example, we know that the a data-set which is not biased and where the order of magnitude varies considerably, is expected to follow this law. This point only answers "where" or "when", but neither provides any physical insight (an understanding from the first principle) nor answer "why". Thus, it needs more investigations. Interestingly, this type of investigation does not require sophisticated equipments or software. One can just use a spread-sheet and follow the steps given in Ref. [25] to check the existence of Benford's distribution in other data-set. 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 [27],

Recently, Eq. 3 has been used in Ref. [7] 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 Eq. 3 for a set of physical properties. Another version of Benford's law has been introduced in [29], and the authors have referred to it as strong Benford's law, which provides the probability of a specific number "s" in a data-set (see definition 1.4.3 in [29]). One can also check validity of strong Benford's law for a data-set using a spread-sheet. 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 [15], which may be viewed as a noise introduced by a person or a group of person 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 [23] 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 the similar directions and those investigations would provide more physical insights on: Why does Benford's law work universally?

Acknowledgment:

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.

[1] S. Newcomb, A. J. of Math. **4**, 39 (1881).
[2] F. Benford, Proc. of A. phil. Society **78**, 551 (1938).
[3] A. Berger, T. P. Hill, *et al.*, Probability Surveys **8**, 1 (2011).
[4] A. Adhikari and B. Sarkar, Sankhyā: The Indian Journal of Statistics, Series B , 47 (1968).
[5] A. Berger, T. P. Hill, and E. Rogers, "Benford online bibliography," `http://www.benfordonline.net/list/chronological` (2009).
[6] M. A. Moret, V. de Senna, M. G. Pereira, and G. F. Zebende, International Journal of Modern Physics C **17**, 1597 (2006).
[7] T. Alexopoulos and S. Leontsinis, Journal of Astrophysics and Astronomy **35**, 639 (2014).
[8] M. Sambridge, H. Tkalčić, and A. Jackson, Geophysical research letters **37** (2010).
[9] B. Busta and R. Weinberg, Managerial Auditing Journal **13**, 356 (1998).
[10] J. L. H. Cáceres, J. L. P. García, C. Martínez Ortiz, and L. G. Domínguez, Electronic Journal of Biomedicine **1**, 27 (2008).
[11] S. Docampo, M. del Mar Trigo, M. J. Aira, B. Cabezudo, and A. Flores-Moya, Aerobiologia **25**, 275 (2009).
[12] G. Sottili, D. M. Palladino, B. Giaccio, and P. Messina, Mathematical Geosciences **44**, 619 (2012).
[13] M. J. De Ceuster, G. Dhaene, and T. Schatteman, Journal of Empirical Finance **5**, 263 (1998).
[14] C. Durtschi, W. Hillison, and C. Pacini, Journal of forensic accounting **5**, 17 (2004).
[15] C. Bruce Busta and R. Sundheim, Center for Business Research **95**, 106 (1992).
[16] J. Deckert, M. Myagkov, and P. C. Ordeshook, Caltech/MIT Voting Technology Project Working Paper (2010).
[17] A. S. De and U. Sen, EPL (Europhysics Letters) **95**, 50008 (2011).
[18] A. D. Rane, U. Mishra, A. Biswas, A. Sen, U. Sen, *et al.*, Physical Review E **90**, 022144 (2014).

[19] G. Bhole, A. Shukla, and T. Mahesh, Chemical Physics Letters **639**, 36 (2015).
[20] L. Shao and B.-Q. Ma, Physica A: Statistical Mechanics and its Applications **389**, 3109 (2010).
[21] I. Ltkebohle, "Exoplanet data archeive for confirm planets," http://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=planets (2016).
[22] P. Pintr, V. Peřinová, A. Lukš, and A. Pathak, Planetary and Space Science **75**, 37 (2013).
[23] P. Pintr, V. Peřinová, A. Lukš, and A. Pathak, Planetary and Space Science **99**, 1 (2014).
[24] "Looking for habitable planets beyond solar system," http://www.natureasia.com/en/nindia/article/10.1038/nindia.2014.123 (2014).
[25] I. Ltkebohle, "Step by step instructions for using benford law," http://www.theiia.org/intAuditor/media/files/Step-by-step_Instructions_for_Using_Benford's_Law[1].pdf.
[26] G. Bhole, A. Shukla, and T. Mahesh, arXiv preprint arXiv:1406.7077 (2014).
[27] T. P. Hill, Proceedings of the American Mathematical Society **123**, 887 (1995).
[28] "Effective teamperature," https://en.wikipedia.org/wiki/Effective_temperature.
[29] S. J. Miller, A. Berger, and T. Hill, "The theory and applications of benfords law," (2015).

AP Statistics
Project 3

This project focuses on a scientific paper titled *Benford's distribution in extrasolar world: Do the exoplanets follow Benford's distribution?* by Abhishek Shukla, Ankit Kumar Pandey, and Anirban Pathak. The paper was found on the Cornell University Library website[1] and was published on June 13, 2016, roughly a month after NASA's *Kepler* mission announced its discovery of 1284 exoplanets.[2,3]

Below is a list of the *statistical terms* found in the paper. They are listed in paragraph form roughly in the order found in the paper. Following each term is a rigorous definition found on Wolfram MathWorld.[4] An interpretation of each term in the context of the paper is also provided. Finally, if we studied any of the terms in AP Statistics, an attempt has been made to describe the term in greater mathematical, topical, and conceptual depth.

> A *distribution*, in general, is a "description of the relative numbers of times each possible outcome will occur in a number of trials."[5] In the paper, the distribution of several parameters that describe the exoplanets are examined: radial velocity, orbital period, mass, volume, density, orbital semi-major axis, radius, stellar age, stellar distance, temperature, total proper motion, and longitude. To quantify the distributions of these many parameters, the authors use "the function describing the probability that a given value will occur...called the probability density function (abbreviated PDF)."[6] PDF applies to many of the distributions we have studied in AP Statistics, and when we were introduced to this concept, we sometimes confused PDF with the cumulative density function (CDF): I therefore see it important to make the distinction between PDF and CDF clear in context of the paper. CDF is the function describing the probability that an accumulated set of given values will occur, and a quick look at the graphs on page 4 of the paper shows that the distribution described in this paper is a PDF: If the distribution were a CDF, then the sum of the probabilities of each discrete variable would be greater than one ($P(1) + P(2) + P(3) + ... + P(9) \geq 1$), which is impossible. Indeed, it is nonsensical for multiple digits to all be the first digit of the same number: All events of any digit being the first digit of a number are mutually exclusive. The results in presented in the paper therefore deal with PDFs, not CDFs.
>
> Benford's distribution is the distribution assumed by *Benford's Law*, which states that "in listings, tables of statistics, etc., the digit 1 tends to occur with probability ~30%, much greater than the expected 11.1%,"[7] and that each subsequent digit (2 through 9) tends to occur with a diminishing probability.[8] Mathematically,[9]

---

[1] https://arxiv.org/abs/1606.05678 The article is available on several websites, including that of the Harvard Research Database and Springer Nature.
[2] Exoplanets are formally known as "extrasolar planets." They are the planets of any star besides Sun.
[3] For a complete, up-to-date list of all confirmed exoplanets, see http://exoplanet.eu/catalog/ and https://exoplanets.nasa.gov/newworldsatlas.
[4] http://mathworld.wolfram.com
[5] http://mathworld.wolfram.com/StatisticalDistribution.html
[6] Ibid.
[7] http://mathworld.wolfram.com/BenfordsLaw.html
[8] Because Benford's Law is often presented as this purely mathematical abstraction, it helps to have physical evidence to suggest that the Law is actually valid. We see compelling physical evidence for Benford's Law in the old-fashioned tables of logarithms, in which we "[note] that the first pages are much more worn and smudged than later pages."[7]

Let P(x) represent the probability that the digit x is the first digit of a number. Since P(x) is a "universal probability distribution" over all x, "then it must be invariant under a change of scale."[10] That is, any horizontal dilation of P can be factored out of P(x) into a separate function f(x).

$$P(kx) = f(x) * P(x)$$
$$P(x) = \frac{1}{f(k)} * P(kx)$$
$$\int P(x) * dx = \int \frac{1}{f(k)} * P(kx) * dx$$

Since $\frac{1}{f(k)}$ is a constant, it can be factored out of the integral:

$$\int P(x) * dx = \frac{1}{f(k)} * \int P(kx) * dx$$

Now, consider the CDF of P(x) for all x, the sum of the probabilities of each digit x being the first digit of a number of probability. Since this sum includes the entire sample space of P, it equals 1.[11] Using integral notation,

$$\int P(x) * dx = 1$$

Substituting this result into the previous equation,

$$1 = \frac{1}{f(k)} * \int P(kx) * dx$$

Multiplying by f(k) on both sides,

$$f(k) = \int P(kx) * dx$$

By algebraic manipulation,

$$f(k) = \int P(k * x) * \left(\frac{1}{k}\right) * k * dx$$

Let $u = k * x$. Then $du = k * dx$. By u-substitution,

$$f(k) = \frac{1}{k}\int P(u)du$$

Since $\int P(x) * dx = 1$,

$$f(k) = k^{-1}$$

Now consider that $f'(k) = -k^{-2}$. Setting $k = 1$, $f'(1) = -1$.
Also consider that since $P(kx) = f(x) * P(x)$, by the chain rule,

$$kP'(kx) = f(x) * P'(x)$$

Dividing by f(k) on both sides,

$$\frac{k}{f(k)} * P'(kx) = P'(x)$$

Multiplying by x on both sides, and letting $k = 1$.

$$\frac{x}{f(1)} * P'(x) = x * P'(x)$$

Ms. Sutcliffe and I thought and thought about this next step, but with the sparse work provided by Wolfram MathWorld, we couldn't figure out how to get from the

[9] Thanks to Ms. Sutcliffe for helping me through this math.
[10] Ibid.
[11] This is a source of contention because no indefinite integral can equal a definite constant.

previous step to the next one. Some unexplained property of f(k) is such that $\frac{1}{f(1)} = -x$.

Assuming this is true, and remembering that we found f'(1) = -1,

$$-x * P'(x) = -P(x)$$

Rewriting P(x) as y and P'(x) as $\frac{dy}{dx}$,

$$x * \frac{dy}{dx} = -y$$

Solving the separable differential equation,

$$\frac{1}{y}dy = -\frac{1}{x}dx$$

Integrating, where C is a constant,

$$\int \frac{1}{y}dy = -\int \frac{1}{x}dx$$

$$\ln|y| = -\ln(|x|) + C$$

Adding $\ln|x|$ to both sides,

$$\ln|x| + \ln|y| = C$$

By the sum of logs property,

$$\ln|xy| = C$$

Exponentiating,

$$e^C = |xy|$$

Letting $k = |xy|$, and considering that $|xy|$ will always be positive since Benford's distribution lies in the 1st quadrant,

$$k = xy$$

Dividing by x on both sides,

$$\frac{1}{x} * k = y = P(x)$$

Letting $k = 1$,

$$P(x) = \frac{1}{x}$$

Thus, the probability of selecting a digit as the first digit in a number is correlated with the reciprocal of that very digit.[12] The graphs on page 4 of the paper accordingly reveal the distinctly hyperbolic shape of the $\frac{1}{x}$ function. It is also interesting to note that Benford's distribution "is not a proper probability distribution (since it diverges)."[13] A fundamental rule of probability is that the entire sample space occupies the probability of 1. Indeed, the Normal distribution is a member of the $\frac{1}{1+x^2}$ family of functions[14], a family whose area converges and is therefore a proper probability distribution. The following integral test for convergence/divergence shows that Normal CDF converges and that the distribution is proper:

---

[12] "While Benford's law unquestionably applies to many situations in the real world, a satisfactory explanation has been given only recently through the work of Hill (1998)." Ibid. While Benford empirically derived his law in 1938, much of the details regarding the precise shape of the distribution and how it varies from case to case is the result of modern mathematical inquiry.

[13] Ibid.

[14] Here is yet another source of contention. While the Normal curve resembles $\frac{1}{1+x^2}$, it may actually be defined by $e^{-\frac{x^2}{2}}$. This function cannot be simply integrated and must be treated as the infinite geometric series $\sum_{n=1}^{\infty} \frac{(-1)^{n+1} * x^n}{2^{n-1} * (n+1)!} = 1 - \frac{x^2}{2} + \frac{x^4}{4*2!} - \frac{x^8}{8*3!} + \cdots$. Given that the Normal distribution is proper, $\int_1^{\infty} \frac{(-1)^{n+1} * x^n}{2^{n-1} * (n+1)!}$ should converge, but I have not carried out this test, and verification is needed.

We would like to determine whether the summation below converges or diverges.

$$\sum_{n=1}^{\infty} \frac{1}{n^2+1}$$

The integral test states that

i) If $\int_1^{\infty} f(x)dx$ is convergent, then $\sum_{n=1}^{\infty} f(n)$ is convergent.

ii) If $\int_1^{\infty} f(x)dx$ is divergent, then $\sum_{n=1}^{\infty} f(n)$ is divergent.

Let $f(x) = \frac{1}{x^2+1}$

$$\int_1^{\infty} f(x)dx = \int_1^{\infty} \frac{1}{x^2+1}dx$$

Since we are evaluating to an upper limit of ∞, we must treat the integral above as an improper integral.

$$\int_1^{\infty} \frac{1}{x^2+1}dx = \lim_{a\to\infty} \int_1^{a} \frac{1}{x^2+1}dx$$

$$\lim_{a\to\infty} \tan^{-1} x \,|\, 1 \to a$$

$$\lim_{a\to\infty} \tan^{-1} a - \tan^{-1} 1$$

Evaluating the limit,

$$\frac{\pi}{2} - \frac{\pi}{4} = \frac{\pi}{4}$$

By condition (i) of the integral test, we conclude that the family of functions defined by $\sum_{n=1}^{\infty} \frac{1}{n^2+1}$ converge. Because the Normal distribution belongs to this family, Normal CDF converges and is classified as a proper probability distribution.
In contrast, we would like to determine whether this summation, the harmonic series converges or diverges.

$$\sum_{n=1}^{\infty} \frac{1}{n}$$

Let $f(x) = \frac{1}{x}$

$$\int_1^{\infty} f(x)dx = \int_1^{\infty} \frac{1}{x}dx$$

We are once again faced with an improper integral.

$$\int_1^{\infty} \frac{1}{x}dx = \lim_{a\to\infty} \int_1^{a} \frac{1}{x}dx$$

$$\lim_{a\to\infty} \ln x \,|\, 1 \to a$$

$$\lim_{a\to\infty} \ln a - \ln 1$$

Evaluating the limit,

$$\infty - 0 = \infty$$

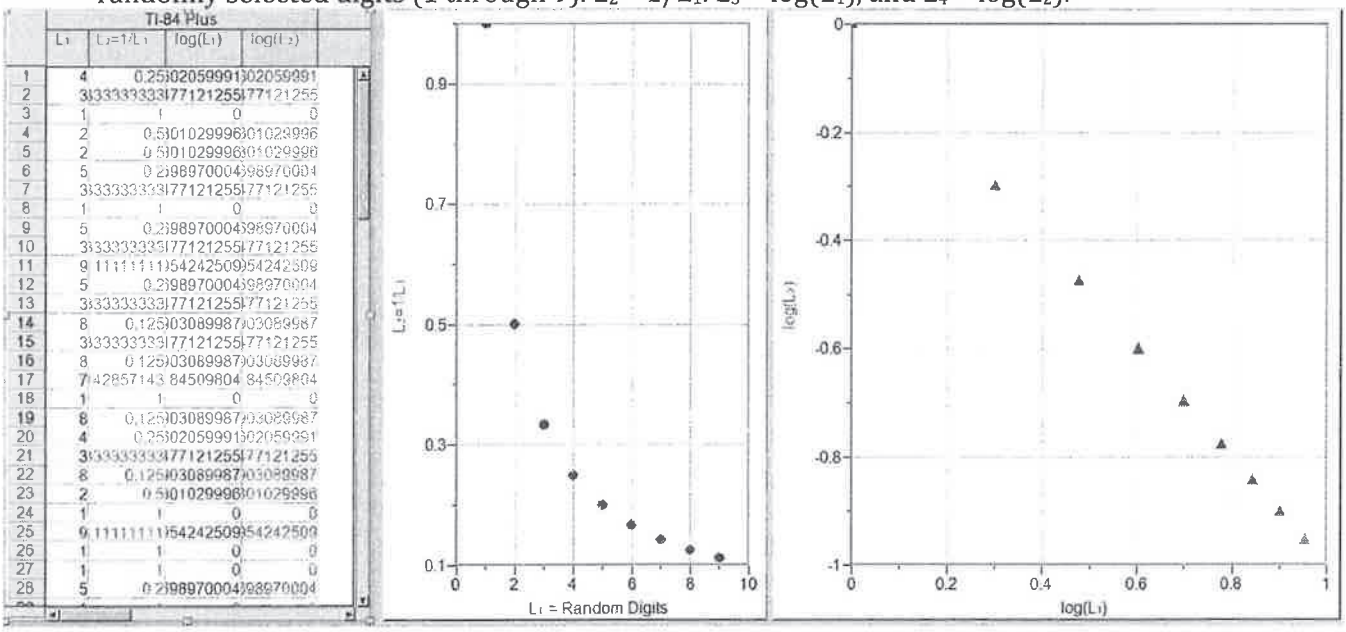
By condition (ii) of the integral test, we conclude that the harmonic series diverges. Since Benford's Law is based on this series, it, too, diverges. That is, the CDF of the distribution is infinite, violating a fundamental law of probability. However, "both the laws of physics and human convention[15] impose cutoffs"[16] on the distribution,

[15] Our textbook includes several problems that use Benford's Law, several of which involve "cases of tax fraud and election fraud"[1] in which a hypothesis test is to be conducted at a .05 = α significance level (for example, see 11.2 #11). Firstly, the fact we conduct a *test* rather than an *estimate* shows that the technical precision of Benford's Law becomes superfluous in answering a yes-or-no question. Secondly, any errors due to treating

and for all intensive purposes, these cutoffs allow for approximations with negligible error.

Later in the abstract, the authors write, "Benford *goodness parameters* [BGPs] are computed to provide a quantitative measure of coincidence of real data with the ideal values obtained from Benford's distribution."[17] In AP Statistics, we studied the Chi-Squared goodness-of-fit test, which involves a quantitative measure of coincidence of real data with ideal values obtained from the Chi-Squared distribution. While the methods to compute these goodness-of-fit parameters are different from distribution to distribution, the concept is the same. The graphs on page 4 of the paper are organized in accordance to descending BGP values, which correspond to descending semblance to Benford's distribution.

In the introduction of the paper, the authors mentioned how Simon Newcomb, "while going through *the logarithms of an unbiased data-set*,"[18] noticed an anomalous behavior in the distribution of digits. The very process by which this distribution was first empirically discovered involves a concept our AP Statistics class covered in 12.2: the linearization of power functions by taking the logarithm of both x and y values. Since Benford's distribution derives itself from $y = \frac{1}{x}$, which can also be written as $y = x^{-1}$, this function can be treated a power function such that $(\log(x^{-1}), \log y)$ is linear. The following illustration demonstrates the validity of Newcomb's observations. In the following data table, $L_1 = 100$ randomly selected digits (1 through 9). $L_2 = 1/L_1$. $L_3 = \log(L_1)$, and $L_4 = \log(L_2)$.



If the digits randomly selected in L_1 followed Benford's distribution, there would be more ones than twos, twos than threes, threes than fours, and so on. Thus, the hyperbola would demonstrate a higher concentration of points to the left, perhaps in the interval (0,2). However, since the hyperbola "thins" out in this interval, this higher concentration would be more discernable in a linearized graph.

Benford's distribution as a proper probability distribution pale in comparison to the relatively large significance level, small sample size, and the variation in the data due to outliers.

[16] Ibid.

[17] See footnote 1.

[18] Ibid.

On page 2 of the paper, the authors go on to say that the "[s]ize 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 ≈ 3300... [w]hich is a reasonable size for statistical analysis of the data-set in general, and for investigation on the existence of Benford's distribution, in particular."[19] In stating the approximate number of detected and confirmed exoplanets, the authors are checking what we call the *Independence condition* in AP Statistics. We check this condition to ensure that our sample belongs to a considerably larger population (at least 10x larger) so that the selection of samples is as subject as possible to randomness, which is fundamental to inference. We also check the Independence condition to ensure that no observation affects any other observation (Imagine testing the water quality in bathtub with gallon-sized samples. If we take too many samples, we end up changing the entire bathtub's qualities [i.e. the concentrations of denser particles that sink to the tub], which makes inference invalid), but since our observations of planets around stars hundreds of lightyears away does not affect those stars and planets in any way,[20] we can ignore this clause of the Independence condition.

On page 3, the authors call Benford's distribution "scale-independent." Wikipedia defines scale-invariance as "a feature of objects or laws that do not change if scales of length, energy, or other variables, are multiplied by a common factor. The technical term for this transformation is a dilatation (also known as dilation), and the dilatations can also form part of a larger conformal symmetry."[21] This fact explains why, when Benford's distribution was derived above, $P(kx) = f(x) * P(x)$. This property contrasts our study of multiplying a random variable by a constant, in which all constituent elements are multiplied by that constant. Benford's distribution does NOT behave this way and is therefore analogous to the *standard deviation* of a random variable: No matter by what factor the random variable is multiplied, the standard deviation remains constant.

Lastly, on page 3, the authors define *significance levels* so that they can categorize the Benford Goodness Parameter (BGP) for each category of data: "BGP values greater than 80 correspond to a good agreement, BGP values in the range (60<BGP≤80) correspond to moderate or intermediate agreement, and BGP values ≤60 implies that data-set don't follow Benford's distribution."[22] While we in AP Statistics define the a result to be statistically significant "if the *P*-value is smaller than α,"[23] a definition that lends itself to a binary situation (in which a result is statistically significant or not) the authors of the paper define their results to have three different degrees of significance. BGPs are also measured differently than *P*-values.

So, what exactly was the paper about? As well-trained, skeptical scientists should, the authors sought to determine which parameters observed by *Kepler* follow Benford's distribution, a distribution that helps investigators detect fabrications. While we know the *Kepler* team does not fabricate its data, the authors' performance of significance tests on the parameters investigates not

---

[19] Ibid.

[20] Please treat Heisenberg Uncertainty Principle as negligible source of interference between our observations and reality.

[21] https://en.wikipedia.org/wiki/Scale_invariance

[22] See footnote 1.

[23] *The Practice of Statistics, 4th ed.* Starnes, Yates, and Moore. W.H. Freeman and Company, 2012. Print. 535.

only the validity of *Kepler*'s fine instruments,[24] but also the nature of extrasolar planets given that the *Kepler* data is precise and accurate. To carry out the test, we assume the authors used the data of at the most 3300/10 = 330 exoplanets (the 10% condition for independence). By calculating BGPs and categorizing the BGPs into predetermined significance levels, the authors found that the mass, volume, density, orbital semi-major axis, orbital period, and radial velocity of the exoplanets very well match Benford's distribution, the total proper motion, stellar age, and stellar distance correspond roughly to Benford's distribution, and that the longitude, radius, and temperature correspond poorly with Benford's distribution. Since nine of the twelve (75%) parameters matched Benford's distribution, the authors concluded that their study yields "a clear affirmative"[25] to the question their paper's title raises.

Was the paper prepared accurately, and would I make any modifications? The authors used the NASA's New Worlds Atlas and EU's Exoplanet Encyclopedia,[26] and my personal experience with those databases (particularly Exoplanet Encyclopedia) has been positive. While I would like to doubt that the authors made any flaws with their information gathering, our practice AP Statistics practice of stating that we have randomly selected our samples makes me skeptical: The authors never mentioned that the samples were randomly chosen. On the reporting standpoint, the authors' mathematical presentation of their was very clear. The qualitative summary that follows however, should have pointed itself to finding which parameters follow Benford's distribution and which parameters do not. To answer the question "Do exoplanets follow Benford's distribution" with "a clear affirmative"[27] sheds no light on the stratification of BGPs found by the study. At the same time, the authors could argue that they solely wanted to deal which *whether* the data follows Benford's distribution, not *why*. Perhaps the question of *why* the distributions of longitude, radius, and temperature fail to follow Benford's distribution should be taken up by another group familiar with the instruments on board *Kepler*. On another note, because the data strictly comes from the Kepler Space Telescope, the title of the paper should technically be, "Do the exoplanets discovered by *Kepler* follow Benford's distribution?"

My reactions to the results presented in the paper can best be encapsulated by the word "awe." It is awe-inspiring that exoplanets, objects so far removed from our vicinity, are largely governed by the same laws that govern our everyday world. Just as reading this paper made me feel like I had contributed to something larger than myself last summer, it also made me feel connected to something much larger than any effort driven by humans on Earth.

In AP Statistics, we often ask ourselves if we are surprised by results. Considering how the work of Johannes Kepler and Isaac Newton mathematically tie together mass (M), volume (V), density (M/V), orbital period (t), orbital semi-major axis ($r = \sqrt[3]{((\frac{T}{2\pi})^2 * G * M)}$), and radial velocity ($\frac{dr}{dt}$), I was not the least bit surprised to see all of these parameters precisely match Benford's distribution. The other six parameters are harder to measure and make sense to not fit Benford's distribution as well. Total proper motion, for example, is most precisely measured when the motion of the exoplanetary solar system only moves in any component of direction other than in our line-of-sight, but this is rarely the case. Stellar age and distance can only be found via modeling (with a few exceptions [The presence of globular clusters, Cepheid variables, etc. allow for more exact measurements]) and are inherently subject to large error. Finally, considering that the hottest (high

[24] According to a NASA spokesperson interviewed on PBS, the instruments aboard *Kepler* are so sensitive that they can detect the light of a firefly in New York all the way from Los Angeles.
[25] See footnote 1.
[26] See footnote 3.
[27] See footnote 1.

temperature) and largest (large radius) exoplanets (called brown dwarves) are the most easily detectable, it was not surprising to see a right skew in both distributions of radius and temperature. As equipment improves, these distributions may begin to resemble Benford's distribution. The diagram below shows a not-to-scale diagram of a few of the previously discovered exoplanets and the skew towards large size (large radius) and proximity to the star (high temperature).

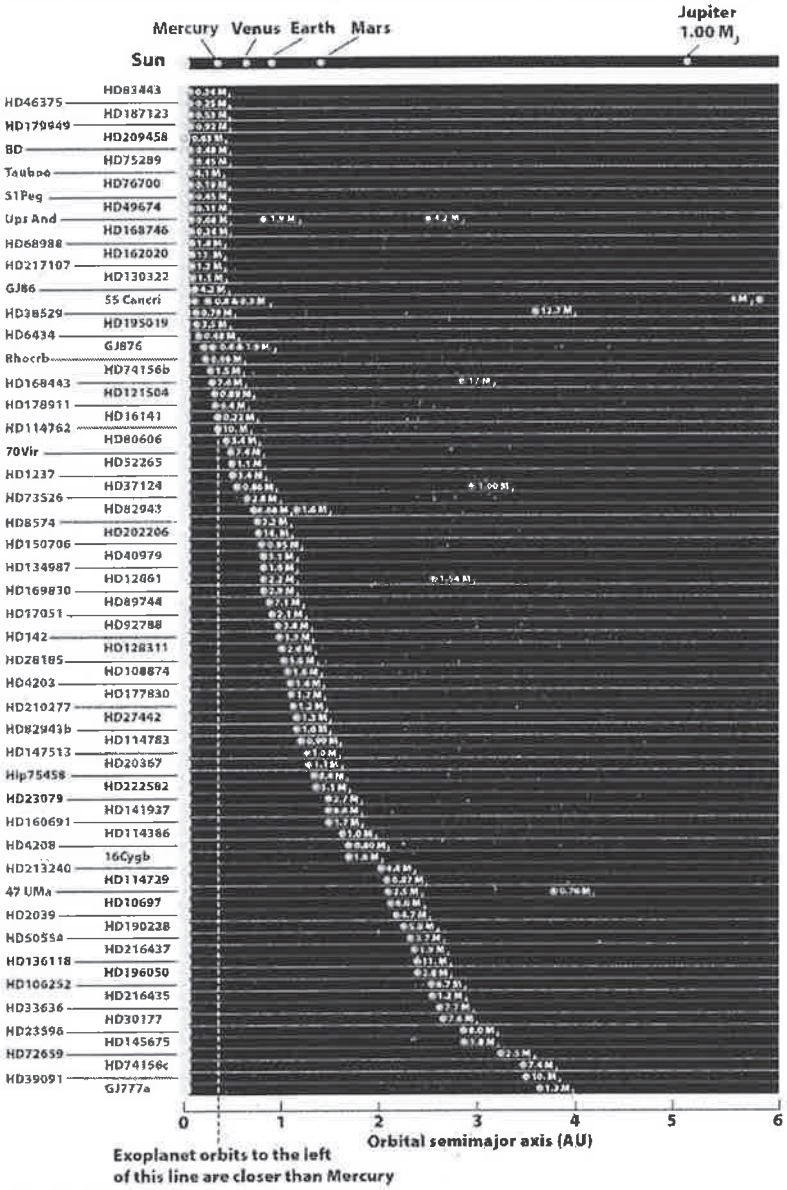


Figure 8-18
Universe, Tenth Edition

Why did I chose this paper as the topic of my final AP Statistics project? Last summer, I had the opportunity to intern with Dr. Mary Urquhart at UT Dallas. A month prior, *Kepler* released the data for 1284 newly discovered exoplanets. Under Dr. Urquhart's direction, I calculated the semimajor axes of the 21 potentially habitable exoplanets, 9 of which belonged to the recently discovered set. $9/21 \approx 43\%$ of the data I used was newly discovered. Given that the exoplanets studied in this paper were randomly selected, then the authors relied on roughly $1284/3300 \approx 39\%$ of the new

data. The proximity of 43% and 39% made me realize that my work last summer was part of a larger effort. Continuing to study exoplanets reminds me of this realization and its associated feelings of unity, pride, and joy.