An extended Benford analysis of exoplanet orbital periods

PATRICK VANOUPLINES [1,*]

[1]University Library – Vrije Universiteit Brussel, 1050 Brussels, Belgium.
[*]Corresponding author. E-mail: Patrick.Vanouplines@vub.be

Abstract. Shukla, Pandey and Pathak (2017) report about their findings of a Benford analysis applied to the physical properties of exoplanets. The present paper gives a short literature overview of previous research on exoplanets. We describe the methods to perform an extended Benford analysis, which considers, both the first digit, and also other digits and digit combinations. Methods for testing conformity with the Benford distribution are discussed and applied to the digits of orbital data of exoplanets.
A first result of this research is that the used data pass most of the tests. It is observed that for most tests on the orbital period values, the almost 4,000 presently known and confirmed exoplanets seem to be sufficient. The analysis of the first, second and third digits (and combinations of these digits) shows a good agreement with the Benford distribution. The analysis of the last two digits indicates that the last significant zero gets lost easily during the export from the exoplanet database. The summation analysis isolates exoplanets with extremely long orbital periods.

Keywords. Benford's law – Benford analysis – exoplanets – orbital period.

1. Introduction

Only a small number of publications discuss the use of Benford's law in astronomy. The paper by Shukla, Pandey and Pathak (2017) on the evaluation of the validity of Benford's law regarding properties of extrasolar planets was the inspiration for a more in-depth study on which we report here. We refer to the previous paper as SPP2017.

1.1 Benford's law and the Benford distribution

Benford's law is not new in scientific literature. Already at the end of the nineteenth century, the astronomer Simon Newcomb noticed that the first pages of books containing logarithm tables where more worn than the other pages. Newcomb (1881) proposed a law expressing the relation between the frequency of numbers starting with a smaller digit and numbers starting with a higher digit: he stated that the probability (as a fraction) of a number starting with digit d (from 1 to 9, the leading zero not being significant) is given by $\log(d + 1) - \log(d)$. He went even further in his paper: he also tabulated the probabilities of the second digit (from 0 to 9). About the third and fourth digits he writes: "In the case of the third figure the probability will be nearly the same for each digit,

and for the fourth and following ones the difference will be inappreciable". The last sentence of Newcomb's paper is "It is curious to remark that this law would enable us to decide whether a large collection of independent numerical results were composed of natural numbers or logarithms".

Newcomb's paper remained unnoticed for a few decades. Frank Benford 1938), apparently unaware of Newcomb's publication, rediscovered the phenomenon (again inspired by the wear and tear of logarithm tables) and tested data sets such as the surface areas of rivers, the sizes of US populations, physical constants, molecular weights, mathematical constants, numbers contained in an issue of Reader's Digest, street addresses of persons listed in American Men of Science and death rates, in total covering more than 20,000 observations. Benford also concludes that "The frequency of first digits thus follows closely the logarithmic relation". When Benford considers the other digits, he writes that also the previous digits should be taken in account. He arrives at formulas describing the distributions of the digits, similar to the equations in the present paper. For many data sets, often referred to as 'natural data', the proportion of the smaller digits is much bigger than the proportion of the larger digits. Counterintuitively, following the Benford distribution, almost fifty per cent of the numbers start with digits 1 and 2. In the present paper we use a consequent notation, not only for the individual digits, but also for combinations of digits.

A Benford analysis looks at the individual digits in numbers. Take, for example, the number 65.4321:

- 6 is the first digit,
- 5 is the second digit,
- 4 is the third digit,
- and so on.

Considering multiple digits:

- the first two digits are 65,
- the second two digits are 54,
- the first three digits are 654,
- the last two digits are 21.

The number 654,321 is, for a Benford analysis, equivalent with 6.54321 and also with 0.0654321.

Besides, we also consider the last two digits, and we perform a summation analysis. These two analyses were introduced in the field of financial auditing, see Nigrini (2012). We call this group of analyses the extended Benford analysis, while we realize that the analysis of the last two digits and the summation analysis are, in fact, not a Benford analysis.

1.2 Previous digit analysis of exoplanet data

The authors of SPP2017 state in their conclusion: "The validity of Benford's law is investigated for the first time for exoplanets.". This is not entirely true. Sambridge, Tkalčić, and Jackson (2010) performed a Benford analysis on the masses of exoplanets, among other physical data sets. Their data set contained only 401 exoplanets, at the time of their investigation. It is interesting to note that in the exoplanet mass data, as the authors describe it, a "bump" occurs. There is an excess of values where the first digit is 6 (9.5%, where the Benford proportion is 6.7%). The authors state that this "difference is subject to both sampling and observational error but would correspond to an excess of 11 planets being erroneously assigned a mass with first digit 6.". Hair (2014) reports about his study of exoplanet masses, with 758 confirmed exoplanets and 3455 Kepler candidate exoplanets (contained in the data set in November 2013). There remains a bump in the exoplanet mass data, although smaller (8.6%) for the first digit 6. Kossovsky (2012) describes his investigation on a dataset of early September 2012, with 834 exoplanets, and he discusses the same subject later in somewhat more detail (Kossovsky, 2015, p 34-35, and p 132). This author gives Benford analyses for exoplanets' mass, angular distance, semi-major axis size, orbital eccentricity, and orbital period. His results are summarized in Figure 1.

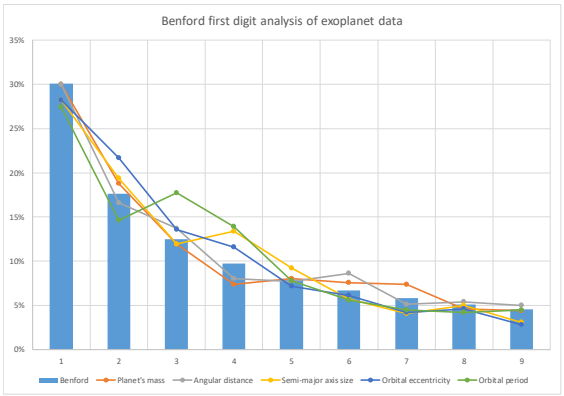

Figure 1. Graphical representation of Kossovsky's Benford first digit analysis of data about the mass,

angular distance, semi-major axis, orbital eccentricity, and orbital period of 834 exoplanets (after Kossovsky 2015, p 35).

2. Extended Benford analysis methods

The authors of SPP2017 write (p 7) that they investigate whether the second most significant digit (for orbital periods of extrasolar planets) also follows Benford's law, but what the authors mean, in fact, corresponds to the first two digits (FTD).

2.1 *Requirements for Benford analyses*

In order to perform a reliable Benford analysis, data should at least: span several magnitudes, consist of thousands of values, not contain fixed nor rounded values. Further descriptions of the requirements for a potential Benford compliant data set can be found in Nigrini (2012, p 21-23) and Miller (2015, p 193).

2.2 *Types of digit analyses*

In this section an overview is given of the Benford probabilities for the most significant digits, and the last two digits. Besides these probabilities, also the summation analysis is described; this analysis is often used in accounting and fraud detection.

To express the Benford probabilities we use the notation of Kossovsky (2015). In this notation, the probability of the first digit (FD) *d* is calculated by

$$Probability\ [1st\ digit\ is\ d] = Log_{10}(1 + 1/d) \quad (1)$$

For the first two digits (FTD) and the first three digits (F3D) the equations are similar:

$$Probability\left[\begin{matrix}1st\ digit\ is\ p\\ AND\ 2nd\ digit\ is\ q\end{matrix}\right] = Log_{10}(1 + 1/pq) \quad (2)$$

and

$$Probability\left[\begin{matrix}1st\ digit\ is\ p\\ AND\ 2nd\ digit\ is\ q\\ AND\ 3rd\ digit\ is\ r\end{matrix}\right] = Log_{10}(1 + 1/pqr) \quad (3)$$

Note that *pq* does not mean the multiplication *p***q*, but the sequence [digit *p*, followed by digit *q*]. Likewise, *pqr* can be generated by (*p**100)+(*q**10)+(*r*). Notice also that in the three equations above *p*≥1, *q*≥0, and *r*≥0.

The form of the equations changes when the probabilities are calculated for digits that do not include the first digit (i.e. when the first digit can take any value 1..9). The second digit (SD) probabilities are calculated with

$$Probability\ [(2nd\ digit = k \mid any\ 1st\ digit)] = \sum \log_{10}(1 + 1/Dk)\,,\ summed\ over\ all\ D\ \in \{1,2,3,4,5,6,7,8,9\} \quad (4)$$

Similar, but slightly more complicated is the calculation of the probability of the third digit (while the first and the second digit can take any value 1..9). The third digit (3D) probabilities are calculated with

$$Probability\ [3rd\ digit = m \mid any\ 1st, any\ 2nd\ digit] = \sum\sum \log_{10}(1 + 1/DKm)\,, summed\ over\ all\ D \in \{1,2,3,4,5,6,7,8,9\}\ and\ K \in \{0,1,2,3,4,5,6,7,8,9\} \quad (5)$$

Obviously we may also calculate the probabilities of the second and the third digits (while the first digit can take any value 1..9). Note that calculation of the probability of this combination of digits is not mentioned by Kossovsky (2015). In this paper it is probably the first time that this higher order Benford probability is tested on natural data. The second two digits (S2D) probabilities are calculated with:

$$Probability\ [2nd\ digit = p\ AND\ 3rd\ digit = q \mid any\ 1st\ digit] = \sum \log_{10}(1 + 1/Dpq)\ summed\ over\ all\ D\ \in \{1,2,3,4,5,6,7,8,9\} \quad (6)$$

Calculating even higher order probabilities is almost not relevant because such a higher-order distribution is almost flat and all probabilities are too close to an equal distribution over all bins of digits.

On the other hand, an analysis of the last two digits (LTD) is interesting. Depending on the number of digits, the last two digits will be a mixture of second, third, and higher-order probabilities, so that theoretically the last two digits in the range {00, 01, … 98, 99} should all occur with a probability of 1/100, or 1 percent. Adding such an

analysis may show rounding and measurement errors. The LTD-analysis is regularly applied in accounting audits.

The summation Analysis (SA), also called the summation test, is often used in accounting audits. The test is developed by Mark Nigrini as a tool to detect anomalous single or multiple transactions. The test is especially powerful in finding an anomalous big number of expenses starting with the same two digits, that would otherwise go unnoticed. To execute the test, for all ninety combinations of first two digits {10, 11, … 98, 99} the amounts are summed. The multiple occurrence of a particular digit combination might point to a process or other reason that should be investigated.

All the tests described in this subsection can be easily implemented in a spreadsheet file so that by importing, or even a fast copy and paste operation the analyses are swiftly performed, together with statistical analyses (which are described in the next subsection) and with ready for inspection graphical representations of the analyses.

2.3 *Goodness of fit methods*

In this subsection methods are given to test for compliance or conformity of a set of data with the Benford probabilities as established in the previous subsection.

The Z-test checks the hypothesis that the proportion of data (being (a combination of) digit(s)) obeys Benford's law, using the equation

$$Z = \frac{|P_O - P_B| - \left(\frac{1}{2N}\right)}{\sqrt{\frac{P_B(1-P_B)}{N}}} \qquad (7)$$

where

P_O is the proportion of the (combined) digit(s), as observed,

P_B is the proportion of the (combined) digits, as predicted by Benford's law,

N is the number of observations,

Z is the Z-statistic.

With the value of the Z-statistic we calculate how much the proportions, based on the observations, deviate from the values as predicted by Benford's law. In the figures in section 3, the Z-value of 1.96 is indicated with a green line, and the Z-value of 2.58 is indicated with a red line.

The chi-square test is a general test to check if a set of data is compliant with the Benford distribution. An important drawback of this test is its oversensitivity when larger data sets are used (which is, see the introduction of this paper, a requirement for a decent Benford analysis). See also Kossovsky (2015, p 123) for a discussion at large. Nevertheless, the test is used in the next section, but the results are compared with the outcomes of three other significance tests (Z-test, MAD-test, and SSD test). Notice also that the Z-test is performed on individual digits (or combinations thereof), while the chi-square test, the MAD-test and the SSD-test are more general, reflecting the goodness of fit for all (combinations of) digits at once.

The chi-square statistic is calculated with:

$$\chi^2 = N * \sum_1^{RD} \frac{(P_B - P_O)^2}{P_B} \qquad (8)$$

where

RD = number of relevant digits,

P_B is the proportion of the (combined) digits, as predicted by Benford's law,

P_O is the proportion of the (combined) digit(s), as observed,

N is the number of observations,

χ^2 is the chi-square statistic.

The authors of SPP2017 do not multiply with the number of observations N, which is the reason why they obtain very low values for the chi-square statistic. The consequences are further explored in the next section, when the results for the analysis of the first two digits is performed. When the multiplication with N is done, it is possible to calculate how much the observed values deviate from the values as predicted by Benford's law. For example, the chi-square statistic is calculated as 123.45 and chi-square is 112.02 (for a probability of 5% and 89 degrees of freedom). Since 123.45 is larger than 112.02, the hypothesis can be rejected (the data are not conform to Benford's law).

Nigrini elaborated a test that is less influenced by the number of data. The chi-square test becomes so sensitive to deviations that, as the number of

data increases to 25,000, near perfection is required. Nigrini (2012, p 158) proposed the Mean Absolute Deviation (MAD) test:

$$MAD = \frac{\sum_{i=1}^{K} |P_O - P_B|}{K} \quad (9)$$

where:

K is the number of (combined) digits (for example, 90 for the first two digits),

P_B is the proportion of the (combined) digits, as predicted by Benford's Law,

P_O is the proportion of the (combined) digit(s), as observed.

Table 1. Interpretation of MAD-values as defined by Nigrini (2012, p 160).

First digits (FD)			Conformity
0.000 ≤	MAD	≤ 0.006	Close
0.006 <	MAD	≤ 0.012	Acceptable
0.012 <	MAD	≤ 0.015	Marginal
0.015 <	MAD		None

Second digits (SD)			Conformity
0.000 ≤	MAD	≤ 0.008	Close
0.008 <	MAD	≤ 0.010	Acceptable
0.010 <	MAD	≤ 0.012	Marginal
0.012 <	MAD		None

First two digits (FTD)			Conformity
0.000 ≤	MAD	≤ 0.0012	Close
0.0012 <	MAD	≤ 0.0018	Acceptable
0.0018 <	MAD	≤ 0.0022	Marginal
0.0022 <	MAD		None

First three digits (F3D)			Conformity
0.00000 ≤	MAD	≤ 0.00036	Close
0.00036 <	MAD	≤ 0.00044	Acceptable
0.00044 <	MAD	≤ 0.00050	Marginal
0.00050 <	MAD		None

Other values for interpretation of the MAD-values were published by Nigrini in his earlier papers, but the above table seems to represent the most "definitive" values.

Kossovsky (2015) proposes the Sum Squares Deviation (SSD) to compare the observed proportions and the expected proportions as predicted by Benford's Law. Like Nigrini's MAD value, the SSD is independent of the number of observations. The SSD can be calculated with:

$$SSD = \sum_{i}^{RD} (T_o - T_e)^2 \quad (10)$$

where:

RD is the number of (combined) digits (for example, 90 for the first two digits)

T_o is the observed percentage of numbers (T_o = 100 x P_O)

T_e is the expected percentage of number, as expected by Benford's law (T_e = 100 x P_B).

Notice that in equation (10) percentages are used, according to Kossovsky (2015, p 128): "This is so in order not to deal with extremely small fractional values which are often confusing and very hard to remember". Notice also that RD in Kossovsky' SSD formula is equivalent to K in Nigrini's MAD formula.

As a consequence for not including the number of observations, only a subjective interpretation can be given for the SSD value (just like for the MAD value only some arbitrary cut-off points are proposed by Nigrini). This interpretation is represented in Table 2 for SSD values. While Nigrini gives his MAD critical values for the interpretation of the first three digits (and not for the analysis of the last two digits), Kossovsky gives cut-off points usable in the analysis of the last two digits (and not for the analysis of the first three digits).

Table 2. Kossovsky's SSD arbitrary cut-off points (After Kossovsky 2015, p 133).

First digits (FD)			Conformity
0 ≤	SSD	< 2	≈ Perfectly Benford
2 ≤	SSD	< 25	Acceptably Close
25 ≤	SSD	< 100	Marginally Benford
100 ≤	SSD		Non-Benford

Second digits (SD)			Conformity
0 ≤	SSD	< 2	≈ Perfectly Benford
2 ≤	SSD	< 10	Acceptably Close
10 ≤	SSD	< 50	Marginal
50 ≤	SSD		Non-Benford

First two digits (FTD)			Conformity
0 ≤	SSD	< 2	≈ Perfectly Benford
2 ≤	SSD	< 10	Acceptably Close
10 ≤	SSD	< 50	Marginal
50 ≤	SSD		Non-Benford

Last two digits (LTD)			Conformity
0 ≤	SSD	< 4	≈ Perfectly Benford
4 ≤	SSD	< 40	Acceptably Close
40 ≤	SSD	< 100	Marginal
100 ≤	SSD		Non-Benford

2.4 *Data set*

The data for this research were downloaded on January 5th , 2018 (at 00:10:46 UTC), with a total of 3,490 at that moment confirmed exoplanets, from the NASA ExoplanetArchive, http://exoplanetarchive.ipac.caltech.edu (further on referred to as the NASA database). A month later, on February 5th, there were already 3,605 exoplanets in this archive.

For a more in-depth examination of the last two digits, orbital period data (expressed in days) for at that time 3,605 confirmed exoplanets were downloaded on 21 February 21st, 2018 (at 13:21 UTC) from exoplanet database http://exoplanet.eu/catalog/ (further on referred to as the EU database) because the Comma Separated Values (CSV) output from this interface gives also the significant ending zeroes.

3. Extended analysis of exoplanet orbital periods

3.1 First Digit (FD) analysis

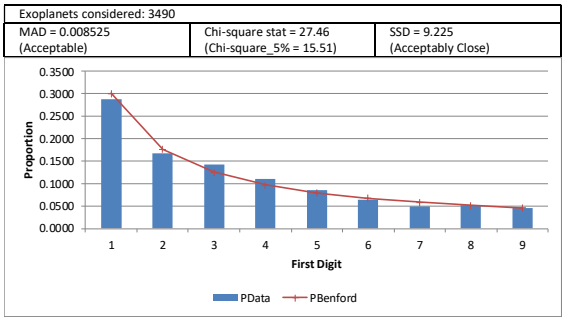

Figure 2. First digit analysis of the orbital period of 3,490 exoplanets.

At first sight, the graph for the analysis of the first digits shows a remarkable correspondence with the Benford distribution. However, the chi-square statistic is, with a value of 27.46, well above 15.51, the 5%-level. Also the MAD-value is considered 'acceptable', which is not a close conformity. Indeed, deviations for most of the digits are clearly visible in the graph. The deviations are very similar to those found in SPP2017, with slightly less orbital periods starting with 1 and 2, slightly more starting with 3, 4, and so on. Multiplied with the number of observations (3,207 exoplanets) used in SPP2017, the value for the chi-square statistic given in SPP2017 is also high (0.011 * 3,207 = 35.28). A few Z-scores, see Fig. 3, are high, especially for first digits 3 and 4. The SSD-value of 9.22 is, compared to the SSD-values for the analyses of the other digits, rather high, but still within the range 'Acceptably Close'.

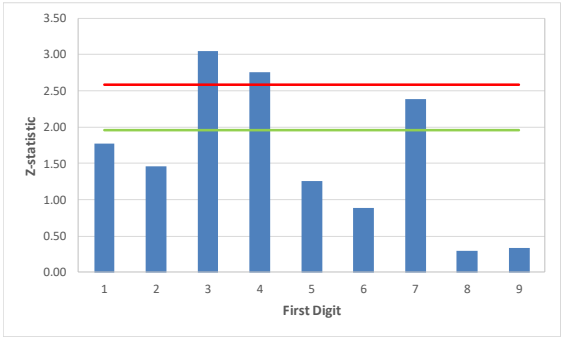

Figure 3. Graphical presentation of the Z-scores for the first digit analysis. Green line is at 1.96, red line is at 2.58.

3.2 Second Digit (SD) analysis

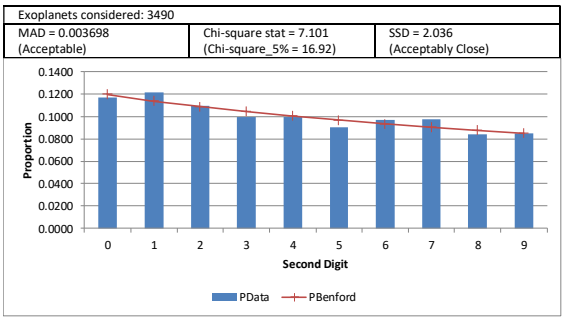

Figure 4. Results of the analysis of the second digits of the orbital period of 3,490 exoplanets.

The analysis of the second digit gives very low values for the chi-square statistic. The Z-scores are low: none of the scores exceeds 1.96. The MAD-value (which can be considered as 'close'), and the SSD-value is almost within the range 'Perfectly Benford'.

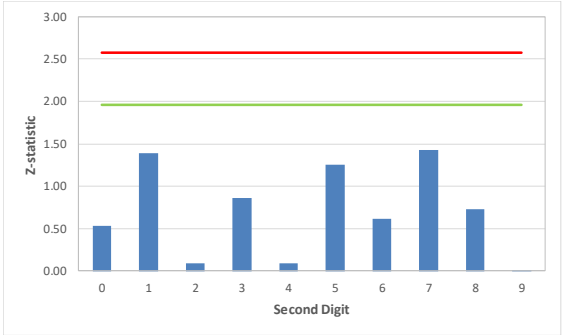

Figure 5. Graphical presentation of the Z-scores for the second digit analysis. Green line is at 1.96, red line is at 2.58.

3.3 Third Digit (3D) analysis

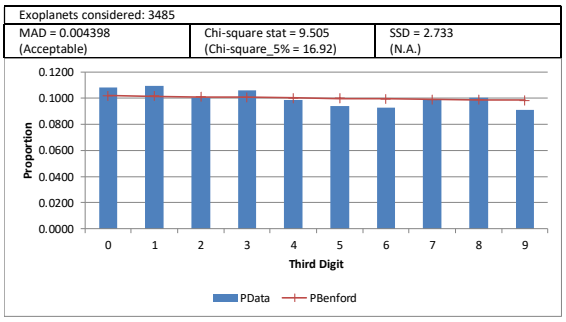

Figure 6. Results of the analysis of the third digits of the orbital period of 3,485 exoplanets.

The analysis of the third digit leads to results that are similar to those of the analysis of the second digit: low values for both the chi-square statistic and the MAD-value indicate a close conformity with the Benford distribution. Notice that the number of exoplanets is smaller (3,485) than in the two previous analyses (3,490): obviously there are five exoplanets whose orbital period is not represented by more than two digits. Again, the Z-scores are low, which confirms the close conformity with the Benford distribution, for all third digits. The SSD-value is also low, but Kossovsky (2015) did not specify cut-off points for the analysis of the third digit analysis, so that an interpretation cannot be given.

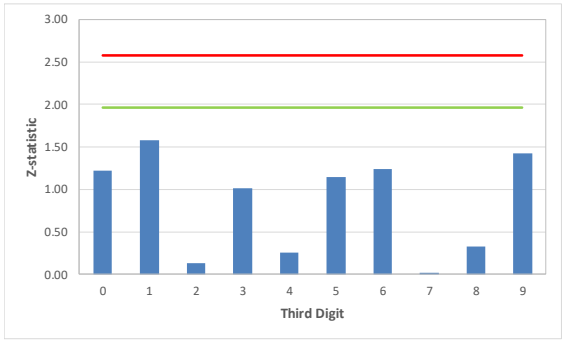

Figure 7. Graphical presentation of the Z-scores for the third digit analysis. Green line is at 1.96, red line is at 2.58.

3.4 First Two Digits (FTD) analysis

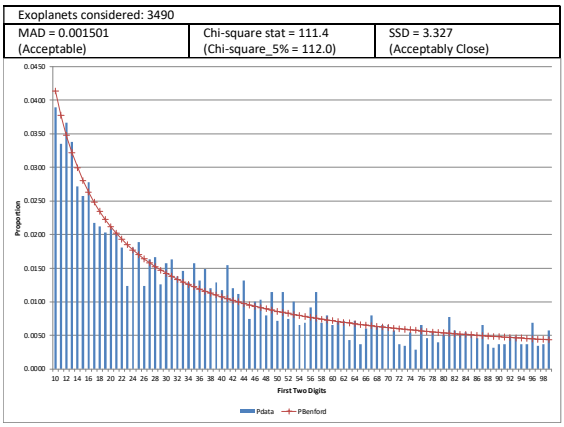

Figure 8. Results of the analysis of the first two digits of the orbital period of 3,490 exoplanets.

The analysis of the first two digits gives a low value for the chi-square statistic. Also the low MAD-value points in the direction of an acceptable conformity with the Benford distribution.

The authors of SPP2017 write that they see a kind of evolution in the values of their chi-square statistic with the number of exoplanets that is being used in the analysis. This can be confirmed with a test where gradually more orbital periods of exoplanets are analysed. The following graph gives the result for 100, 200, … , 3,400, and 3,490 analysed orbital periods of exoplanets.

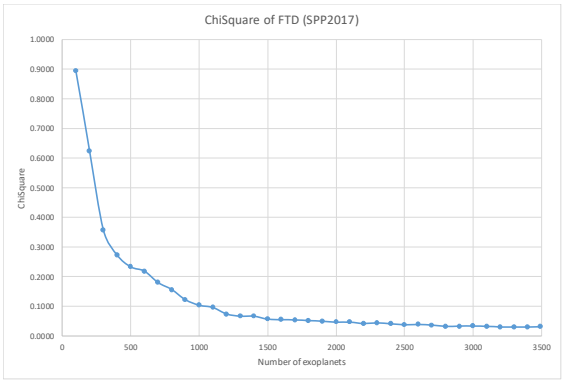

Figure 9. Evolution of the chi-square statistic when performing an analysis of the first two digits of the orbital periods of exoplanets. The number of exoplanets is increased in steps of 100 exoplanets. The equation used to calculate the chi-square statistic is that of SPP2017.

Notice that the chi-square value of 0.06 (for 1898 exoplanets) and 0.036 (for 3207 exoplanets) reported by SPP2017 is confirmed in the above graph. However, when we use equation 8, taking into account the number of observations, then the

results of this analysis are different, as shown in the following graph.

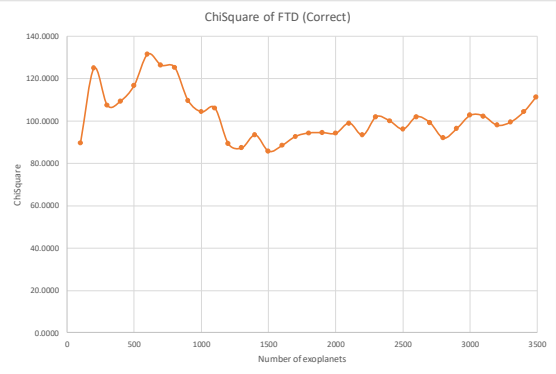

Figure 10. Evolution of the chi-square statistic when performing an analysis of the first two digits of the orbital periods of exoplanets. The number of exoplanets is increased in steps of 100 exoplanets. The equation used to calculate the chi-square statistic is the classical formula (8).

The somewhat larger variations on the left side of the graph in Fig. 10 are of the result of too few observations (*N*). These values for the chi-square statistic are sometimes above 112.02, which means that in these cases (with only few exoplanets)the first two digits of the orbital periods are not conform to the Benford distribution.

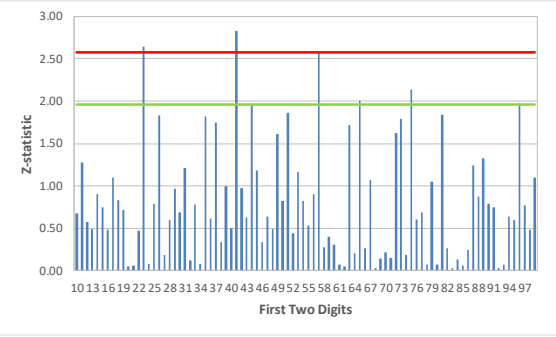

Figure 11. Graphical presentation of the Z-scores for the first two digits analysis. Green line is at 1.96, red line is at 2.58.

The Z-scores are somewhat higher than those of the previous analyses, which does not come as a surprise, because the chi-square values were also a little higher. Still, most of the digits occur in proportions as expected. Also the SSD-value is somewhat high, but still in the conformity range 'Acceptably Close'.

3.5 Second Two Digits (S2D) analysis

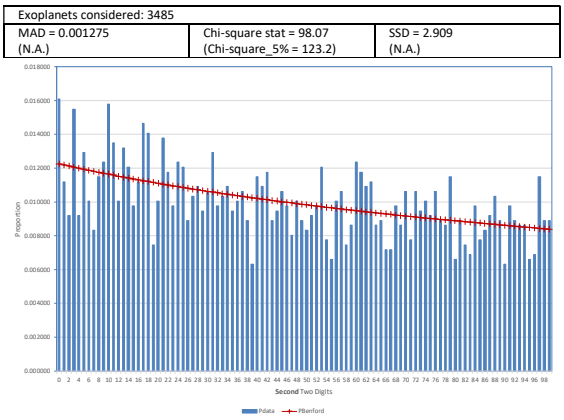

Figure 12. Results of the analysis of the second two digits of the orbital period of 3,490 exoplanets.

At first glance, in Fig. 12, the conformity of the probabilities of the second two digits is not very close. It should be understood that this visual interpretation of the graph is dangerous. Indeed, the difference between the chi-square statistic is rather big (and the value is well below chi-square, for 99 degrees of freedom). The MAD-value seems to be low, but one has to keep in mind that for this type of analysis no MAD-indicator was set up by Nigrini.

The Z-scores for the analysis of the second two digits do not show anything surprising: all scores are low, indicating a high conformity with the Benford distribution. The SSD-value is low, but Kossovsky (2015) did not specify cut-off points for the analysis of the second two digits.

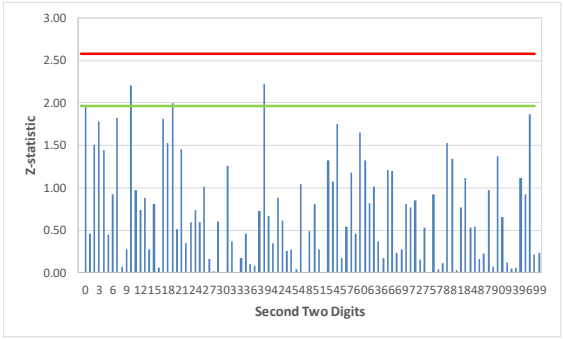

Figure 13. Graphical presentation of the Z-scores for the second two digits analysis. Green line is at 1.96, red line is at 2.58.

3.6 First Three Digits (F3D) analysis

Exoplanets considered: 3485		
MAD = 0.0004293 (Acceptable)	Chi-square stat = 896.9 (Chi-square_5% = 969.9)	SSD = 2.962 (N.A.)

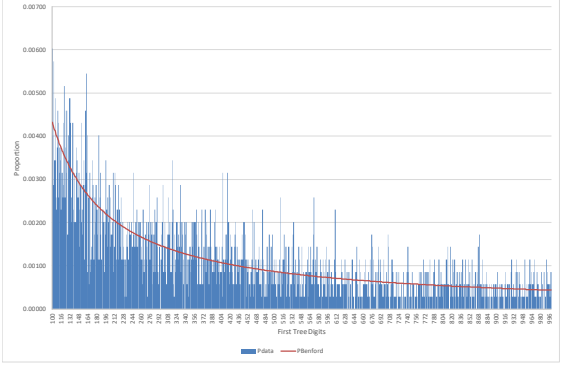

Figure 14. Results of the analysis of the first three digits of the orbital period of 3,485 exoplanets.

The analysis of the first three digits shows an acceptable conformity with the Benford distribution, which is confirmed by the chi-square statistic and the low MAD-value. Some deviations occur, which is also visible in the Z-scores. The SSD-value is low, but Kossovsky (2015) did not specify cut-off points for the analysis of the second two digits.

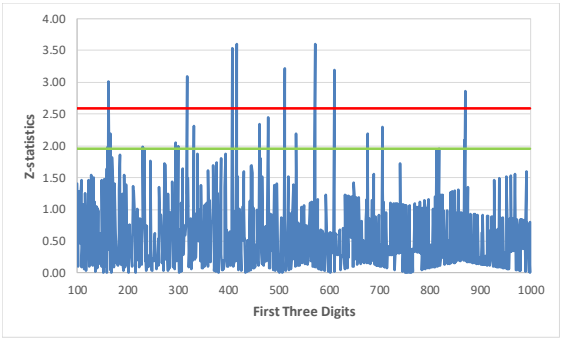

Figure 15. Graphical presentation of the Z-scores for the first three digits analysis. Green line is at 1.96, red line is at 2.58.

3.7 Last Two Digits (LTD) analysis

Exoplanets considered: 3490		
MAD = 0.002425 (N.A.)	Chi-square stat = 425.2 (Chi-square_5% = 123.2)	SSD = 12.18 (N.A.)

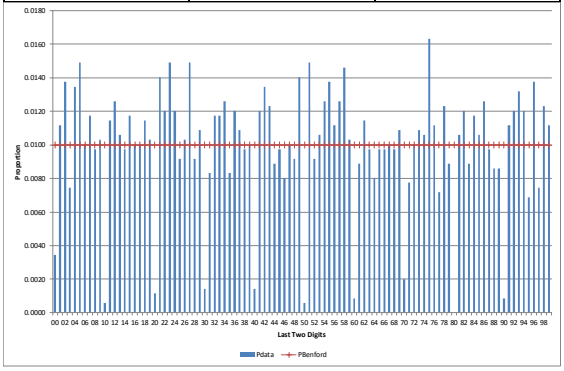

Figure 16. Results of the analysis of the last two digits of the orbital period of 3,490 exoplanets.

The analysis of the last two digits gives a remarkable picture: the tens are not absent, but much less present. A closer look at the process of downloading the data from the database and importing the data into the Excel file (to perform the analyses) shows that significant zeroes get lost, so that it is only when a zero is present before the decimal point that the last two digits become a ten. The problem seems to arise at the moment of export from the database: the CSV output files do not contain the significant ending zeroes, and in other output formats extra decimal zeroes are added. In other words, there is room for improvement regarding the output functionality of the exoplanet database: it would be advisable to foresee an export function that represents all the digits in string format. Using, moreover, the scientific notation adds the advantage that the digits shown will be all the relevant digits, no more, no less.

Some tens are retained, as can be seen in the graph representing the analysis of the last two digits. One example is GJ 328 b for which in the online table the orbital period is given as 4,100±300 days. But a lot of tens are missing, which is also illustrated in the graph representing the results of the Z-test.

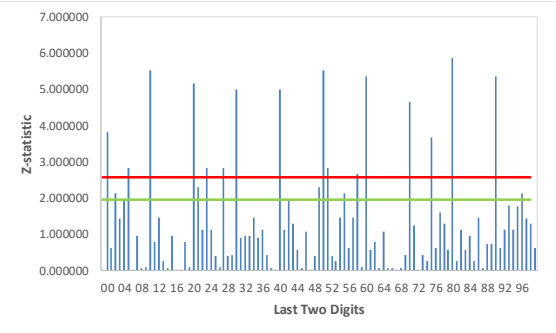

Figure 17. Graphical presentation of the Z-scores for the last two digits analysis. The green line is at 1.96, the red line is at 2.58.

The SSD-value is slightly higher, but Kossovsky (2015) puts the cut-off point for this analysis type higher. Because of the missing tens, however, the SSD-value is without a doubt higher.

We found out that the EU exoplanet database does export the significant ending zeroes to a file of the type CSV. This allowed us to perform an analysis of the last two digits of 3600 planets (in the catalogue on 21 February 2018, 14:21 CET). The results of this analysis are shown in figure 18.

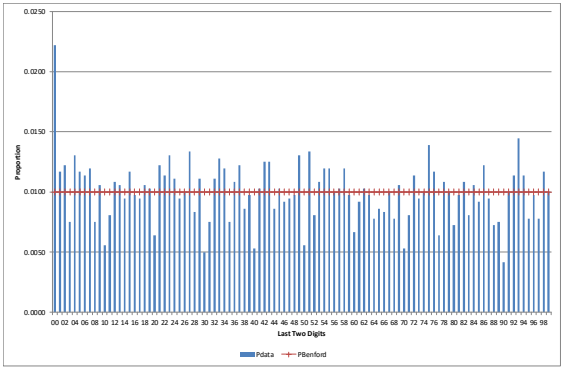

Figure 18. Result of the analysis of the last two digits of the orbital periods of 3,600 exoplanets (data collected from the EU-database on February 21st, 2018, 14:21 CET).

In figure 18 it is clear that there are more exoplanets with an orbital period ending with a zero, but there are still too few exoplanets. The proportion of the exoplanets whose orbital period has 00 as the last two digits is very prominent. These two observations may be explained by a psychological effect, rather than by an astronomical phenomenon.

3.8 Summation analysis (SA)

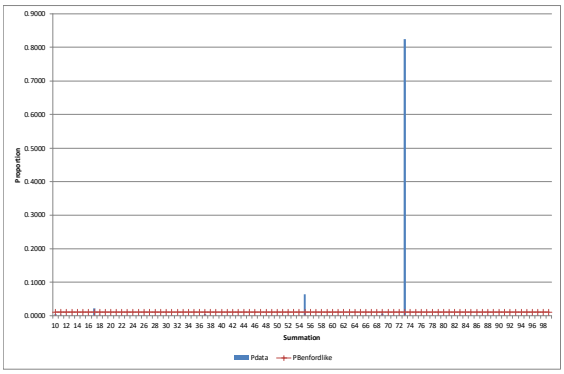

Figure 19. Results of the summation analysis of the orbital period of 3490 exoplanets.

The summation analysis gives an unexpected result: one major peak, and two minor peaks, while all the vertical blue columns were expected to end close to the red line, which is at the bottom of the graph. This result points to an important anomaly, often due to one or more extravagant values. In the present context, this would mean one or more extraordinary (long) orbital periods. The highest peak is at 73 (proportion 0,82). The two other minor peaks are at 55 (proportion 0,065) and 17 (proportion 0,011). It is obvious that these high peaks are generated by long orbital periods of three exoplanets. In Table 3 the eleven exoplanets with the longest orbital periods are summarised.

Table 3. List of eleven exoplanets with the longest orbital periods, based on 3,490 planets, as listed on January 5th, 2018 in the NASA Exoplanet Archive http://exoplanetarchive.ipac.caltech.edu.

pl_hostname	pl_letter	pl_orbper [days]
Oph 11	b	$7300000^{+3650000}_{-3650000}$
Fomalhaut	b	$555530^{+184690}_{-353320}$
HR 8799	b	170000
HR 8799	c	69000
HR 8799	d	37000
HR 8799	e	18000.0
HIP 70849	b	$17337.5^{+15512.5}_{-15512.5}$
47 UMa	d	14002^{+4018}_{-5095}
HD 30177	c	11613^{+1837}_{-1837}
DP Leo	b	$10220.0^{+730.0}_{-730.0}$
CFBDSIR J145829+101343	b	$10037.5^{+2737.5}_{-2737.5}$

In a first attempt, the three planets, listed at the top of Table 3, are removed (see Fig. 20).

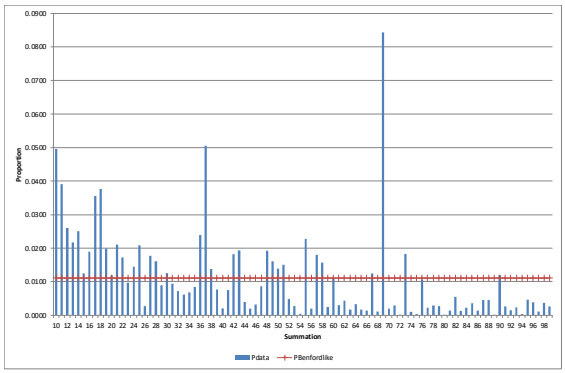

Figure 20. Summation analysis after removal of the three exoplanets with the longest orbital periods.

The graph generated by the summation analysis still shows peaks, even after the removal of the three exoplanets with the longest orbital periods. After the removal of the eleven exoplanets with the longest orbital periods, the graph conforms more with expectation (see Fig. 21).

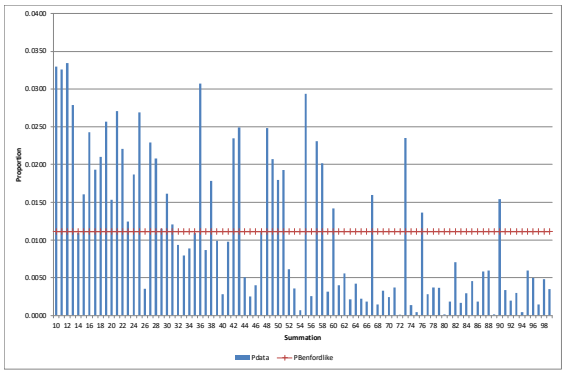

Figure 21. Summation analysis after removal of the eleven exoplanets with the longest orbital periods.

Surprisingly, the value of the orbital period of Oph 11 b is different by a factor of ten in the NASA database (7,300,000 days) and the EU database (730,000 days). This difference could have been noticed earlier by a database user who sorts the exoplanets by orbital period, but it is the extended Benford analysis, more in particular the additional summation analysis, that shed light on this anomaly. Indeed, when an extended Benford analysis is performed during a financial audit, often pointers towards typos and other mistakes are found.

4. Conclusions and recommendations

In this paper we gave an overview of existing methods to perform an analysis of the digits in values of scientific data, and added another analysis method: that of the analysis of the second two digits. As such, we now have a complete set of digit analysis methods, up to the point where calculating even higher order probabilities is no longer useful.

We investigated several tests that provide an indication of conformity with the Benford distribution. These tests have not been developed exhaustively for all the analysis methods used in the present paper. However, the calculated conformity values are all in good agreement. An overview of the obtained results is shown in Table 4.

Table 4. Overview of the results obtained with the extended Benford analysis of exoplanet orbital periods.

	Chi-square statistic (5%-level)	MAD-value Interpre-tation	SSD-value Interpre-tation
First Digit	27.46 (15.51)	0.008525 Acceptable	9.225 Acceptably Close
Second Digit	7.10 (16.92)	0.003698 Acceptable	2.036 Acceptably Close
Third Digit	9.51 (16.92)	0.004398 Acceptable	2.733 -
First Two Digits	111.4 (112.0)	0.001501 -	3.327 Acceptably Close
Second Two Digits	98.07 (123.2)	0.001275 -	2.909 -
First Three Digits	896.9 (969.9)	0.0004293 -	2.962 -
Last Two Digits	425.2 (123.2)	0.002425 -	12.18 -

The present extended analysis of the digits of orbital periods of exoplanets throws an interesting light on outliers and precision of the data. In this paper it became clear that the orbital periods of exoplanets are a very nice example of 'natural data' that conform the Benford distribution. The exceptions mentioned in earlier work by other authors were studied in detail.

The excess of, or "bump" in, numbers with 6 as the first digit, reported by Sambridge et al. (2010) regarding exoplanet masses, does not seem to occur in the orbital periods. For a forthcoming paper it will be interesting to investigate if this "bump" is present in a data set of many more exoplanets, or whether this "bump" only appears in the first few hundred exoplanets (and possibly

fades out, as more and more exoplanet data are included).

The present extended analysis showed that by exporting data from the NASA database, which contains detailed data about exoplanets, does not allow to get all the information about significant zeroes behind the decimal separator. It is advised to add such a detailed export facility. The interface of the EU database does allow for an export of a CSV file that contains the significant ending zeroes.

An interesting result comes from the summation analysis, where a few exoplanets with an extremely long orbital period become prominent. For a further Benford analysis one might think about splitting the exoplanets into at least two classes, a subclass with shorter orbital periods, and a subclass with very long orbital periods. The limit between these two classes should be an orbital period of 10,000 days (or lower), as indicated by the summation analysis.

This paper does not give an answer to the question that arose regarding the analysis of the first data; there seems less conformity of the first digits to the Benford distribution. This may be some kind of 'temporary effect' that disappears when more data become available. And yet, this lesser conformity contrasts with the very high conformity results of the higher order analyses.

A forthcoming paper should apply the extended analysis methods to all available data about exoplanets, not only on the orbital periods. The analysis should also be repeated at a time when data for even more exoplanets are well documented. The fact that the exoplanet database continually grows encourages us to repeat the investigation on a yearly basis. It will be interesting to repeat the test when data about extragalactic planets become available.

Acknowledgements

This research has made use of the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with the National Aeronautics and Space Administration under the Exoplanet Exploration Program.

Use was also made of The Extrasolar Planets Encyclopaedia, developed and maintained by the exoplanet team, based at the Observatoire de Paris.

No financial support was received for this research.

References

Benford, F. 1938, *Proc. Am. Phil. Soc.*, **78**, 551.

Hair T. 2014, *Benford's Law of First Digits and the Mass of Exoplanets*, Joint Meetings American Mathematical Association and Mathematical Association of America, https://www2.fgcu.edu/CAS/MathBS/files/Hair-ppt-Benford.ppsx

Kossovsky, A.E. 2012, Rev. Cienc. Econ., ***30***, 179.

Kossovsky, A.E. 2015, *Benford's Law: Theory, the General Law of Relative Quantities, and Forensic Fraud Detection Applications*, World Scientific Publishing.

Miller, S.J. (ed.) 2015, *Benford's Law: Theory and Applications*, Princeton University Press.

Newcomb, S. 1881, *Am. J. Math.*, **4**, 39.

Nigrini, M.J. 2012, *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*, John Wiley & Sons.

Sambridge, M., Tkalčić, H., Jackson, A. 2010, *Geophys. Res. Lett.*, ***37***, L2230.

Shukla, A., Pandey, A.K., Pathak, A. 2017, *J. Astrophys. Astr.*, ***38***, 1.