BeyondBenford: An R Package to Determine Which of Benford’s or BDS’s Distributions is the Most Relevant
Stéphane Blondeau da Silva

To cite this version:
Stéphane Blondeau da Silva. BeyondBenford: An R Package to Determine Which of Benford’s or BDS’s Distributions is the Most Relevant. 2019. hal-02310013

HAL Id: hal-02310013
https://hal.archives-ouvertes.fr/hal-02310013
Submitted on 9 Oct 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
BeyondBenford: An R Package to Determine Which of Benford’s or BDS’s Distributions is the Most Relevant
by Stéphane Blondeau Da Silva

Abstract

The package BeyondBenford compares the goodness of fit of Benford’s and Blondeau Da Silva’s (BDS’s) digit distributions in a dataset. The package is used to check whether the data distribution is consistent with theoretical distributions highlighted by Blondeau Da Silva or not: this ideal theoretical distribution must be at least approximately followed by the data for the use of BDS’s model to be well-founded. It also allows to draw histograms of digit distributions, both observed in the dataset and given by the two theoretical approaches. Finally, it proposes to quantify the goodness of fit via Pearson’s chi-squared test.

1 Introduction

Benford’s Law, also called Newcomb-Benford’s Law, is somewhat surprising; indeed, the first digit \(d, d \in [1,9]\), of numbers in many naturally occurring collections of data does not follow a discrete uniform distribution, as might be thought, but a logarithmic distribution (see the recent books of [Mil15] and [BH15]). Discovered by the astronomer Newcomb in 1881 [New81], it was definitively brought to light by the physicist Benford in 1938 [Ben38]. The probability that \(d\) is the first digit of a number is approximately:

\[
\log(1 + \frac{1}{d})
\]

It can also be extended to digits beyond the first one [Hi55], the probability for \(d, d \in [0,9]\), to be the \(p^{th}\) digit of a number being:

\[
\sum_{j=10^{p-2}}^{10^{p-1}-1} \log(1 + \frac{1}{10j+d})
\]

It was quickly admitted that numerous empirical data sets follow Benford’s law: economic data [SHK05], social data [Go15], demographic data [NW95, LSE00], physical data [Kmu69, BK91, NM07, AL14] or biological data [CLR10, FPMAS12] for instance; to such an extent that this law was used to detect possible frauds in lists of socio-economic data [Var72, Nig99, DHP04, Sav06, To09, RGB09] or in scientific publications [AYS14].

Nevertheless many discordant voices brought a significantly different message. By putting aside the distributions known to fully disobey Benford’s law [Rai75, Hi88, TBL00, SF01, Bee09, DMO11], this law often appeared to be a good approximation of the reality, but no more than an approximation [SF01, Sav06, DMO11, GD11, Goo16].

Similar to first digit case, the distributions of digits beyond the first have been observed in various application areas [Gey10, AL14, AYS14] and have also been used to detect frauds [Car88, Tho89, ML106, CG07, Die07, Joe13]. Once more, limits of such methods were emphasized [ML106, CG07, Die07].

Blondeau Da Silva, considering data as realizations of a homogeneous and expanded range of random variables following discrete uniform distributions, showed that, the proportion of each \(d, d \in [0,9]\), as leading digit [BDS18] or as other digit [BDS18] structurally fluctuates. He demonstrated that, in his models, the predominance of 1 as digit (followed by 2 and so on) is all but surprising, and that the observed fluctuations around the values of probability determined by Benford’s Law are also predictable: there is not a single Benford’s Law but numerous distinct laws each of them determined by a parameter, the upper-bound of the considered data.

The huge and growing literature on Benford’s law is available on the online database www.benfordonline.net with well over 1000 papers. Two CRAN Packages already enable to check
whether datasets conform to Benford’s law or not: BenfordTests and benford.analysis. The package BeyondBenford compares the goodness of fit of Benford’s Law, on the one hand, and BDS’s Laws, on the other hand. Indeed, these latter, under certain conditions that we will recall, allow a better reliability of adjustment. The package BeyondBenford calculates the digit distribution in the considered dataset and determines whether it is consistent with BDS’s or Benford’s one. It also provides plotting tools for the visual evaluation of these distributions. We will walk through a detailed example to give an overview of the BeyondBenford package.

2 An example to get familiar with the main functions

Street addresses of Pierre-Buffiere, a small town of approximately 1200 inhabitants in Haute-Vienne (France), are available on:

www.data.gouv.fr/fr/datasets/base-d-adresses-nationale-ouverte-bano/,

which is an open platform for French public data.

After loading the package (library(BeyondBenford)), the code data(address_PierreBuffiere) gives us access to the sample data. This factor contains 346 rows, with each row representing an address number.

3 Are the data consistent with BDS’s model?

This is the first essential question that must be answered in the affirmative. If this is not the case, comparisons are not relevant: the package should not be used. Indeed the use of the package is appropriate when the studied data can be considered as realizations of a homogeneous and expanded range of random variables approximately following discrete uniform distributions. In this model, the data is strictly positive and is upper-bounded, constraint which is often valid in datasets, the physical, biological, demographic, social and economical quantities being limited [BDS19].

Among the different domains studied by Benford [Ben38], some could be well adapted to our model: sizes of populations or street addresses for example (see [BDS19] for a detailed explanation). [JDLR04] advised precisely to use their own similar model in the case of street addresses or when considering the first-page numbers of articles in a bibliography. [BDS18] showed that the model induces a specific distribution of positive integers determined by an upper-bound. Hence, in order to conform as closely as possible to the model, the studied database must have a distribution similar to that described in [BDS18]. In $p^{th}$ digit case, the probability $p_k$ to obtain the number $k \in \left[10^{p-1}; 10^p-1\right]$ (where $u_b$ is the upper-bound) verifies:

$$p_k = \frac{1}{u_b - 10^{p-1} + 1} \sum_{i=k+1}^{u_b+1-10^{p-1}} \frac{1}{i}.$$  

In the studied example, the maximum value of the street number is 74. The associated theoretical distribution for the second digit is plotted in Figure 1.

The BeyondBenford package provides plotting tools to determine whether the data is consistent with BDS’s model: the function dat.distr. The function’s arguments are as follows:

- **dat**: the considered dataset, a data frame containing non-zero real numbers.
- **xlab**: the x-axis label. Default value: xlab="data".
- **ylab**: the y-axis label. Default value: ylab="Frequency".
- **main**: the title of the graph. Default value: main="Distribution of data".
- **theor**: if theor=TRUE BDS’s theoretical distribution is plotted, otherwise only the histogram is represented. Default value: theor=TRUE.
- **nclass**: a strictly positive integer: the number of classes in the histogram. Default value: nclass=50.
- **col**: the color used to fill the bars of the histogram. NULL yields unfilled bars. Default value: col="lightblue".
- **conv**: if conv=1, all values of the dataset are multiplied by $10^k$ where $k$ is the smallest positive integer such that all non-zero numerical values in the newly multiplied data frame have an absolute value greater than or equal to 1. Default value: conv=0.
Figure 1: Theoretical distribution of street numbers in the second digit case, the upper-bound being 74.

- **upbound**: a positive integer, which characterizes the data. All (or most) of the data are lower than this "upper-bound". Default value: upbound=ceiling(max(dat)).
- **dig**: the chosen position of the digit (from the left). Default value: dig=1.
- **col**: the color used to plot BDS’s theoretical distribution. Default value: col="red".
- **ylim**: a two-components vector: the range of y values. Default value: ylim=NULL.
- **border**: the color of the border around the bars. Default value: border="blue".
- **ncchi**: the number of classes for values from $10^{p-1}$ to max(max(data),upbound). If nchi>0, the function returns the chi-squared statistic (with nchi – 1 degrees of freedom) of goodness of fit determined by the different classes. The null hypothesis states that the studied distribution is consistent with the considered theoretical distribution. Default value: nchi=0.
- **legend**: if legend=TRUE, the legend is displayed. Default value: legend=TRUE.
- **bg.leg**: the background color for the legend box. Default value: bg.leg="gray85".

Let us apply the dat.distr function to the address_PierreBuffiere dataset. The output from dat.distr is the data histogram along with optional BDS’s theoretical distributions (Figure 2).

**Example 1.**
```
## Both the histogram and theoretical distribution are represented
dat.distr(address_PierreBuffiere, dig=2, nclass=65)
```

Note that, out of the 346 values, only 217 are taken into consideration here because the values need to have at least 2 digits.

The data distribution looks similar to the one described by BDS’s model (Figure 2). Let us provide a second example in which we numerically determine whether the studied distribution is conform to the theoretical distribution or not:

**Example 2.**
```
## The function returns the chi-squared statistic of goodness of fit determined by nchi classes.

dat.distr(address_PierreBuffiere, dig=2, nchi=4)
```

<table>
<thead>
<tr>
<th>Class freq.</th>
<th>Theor. freq.</th>
<th>Chi2 pval</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;130&quot;</td>
<td>&quot;130.050977504473&quot;</td>
<td>1.08754607580421 0.780081329402347</td>
</tr>
<tr>
<td>&quot;51&quot;</td>
<td>&quot;54.6446927640259&quot;</td>
<td>1.08754607580421 0.780081329402347</td>
</tr>
<tr>
<td>&quot;25&quot;</td>
<td>&quot;26.7009825219254&quot;</td>
<td>1.08754607580421 0.780081329402347</td>
</tr>
</tbody>
</table>

The dat.distr function returns, if requested, the frequencies of each equal-sized class of the dataset and the associated theoretical frequencies. It also returns a data frame containing the value of the chi squared test-statistic and its p-value. Note that the number of classes is limited by the theoretical frequencies that cannot exceed 5 in Pearson’s chi-squared test [Pea00]. In our example, the null hypothesis cannot be rejected: the studied distribution is consistent with the theoretical distribution.
4 Comparisons of the goodness of fit of Benford’s and BDS’s digit distributions in a dataset

4.1 Raw data

The package BeyondBenford provides a function returning the frequencies of each figure at a given position in the considered dataset: obs.numb.dig. Its two arguments dat and dig have already been defined in dat.distr function. The function output is a vector containing the frequencies of each figure in ascending order.

Let us give an example:

Example 3. obs.numb.dig(address_PierreBuffiere, dig=2)

\[
\begin{array}{c}
131242721251721161916
\end{array}
\]

For instance, there are 25 values with the second digit being a 4.

The package BeyondBenford also provides a function returning Benford’s probability that a figure is at a given position: Benf.val. Its argument dig has already been defined in dat.distr function. Its second argument fig stands for the considered figure.

Let us give an example:

Example 4. Benf.val(4, dig=2)

\[
\begin{array}{c}
0.1003082
\end{array}
\]

The package BeyondBenford at last provides a function returning BDS’s probability that a figure is at a given position (once the associated upper-bound has been specified): Blon.val. Its three arguments dig, upbound and fig have already been defined above.

Let us give an example:

Example 5. Blon.val(fig=4, dig=2, upbound=74)

\[
\begin{array}{c}
0.09836642
\end{array}
\]
Table 1: Values of frequency of each figure as second digit in the database, Benford’s and BDS’s theoretical values (the chosen upper-bound being 74). These values are rounded to the nearest ten-thousandth.

<table>
<thead>
<tr>
<th>Figure</th>
<th>Frequency in the database</th>
<th>BDS’s values</th>
<th>Benford’s values</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.1429</td>
<td>0.1436</td>
<td>0.1197</td>
</tr>
<tr>
<td>1</td>
<td>0.1106</td>
<td>0.1247</td>
<td>0.1139</td>
</tr>
<tr>
<td>2</td>
<td>0.1244</td>
<td>0.1136</td>
<td>0.1088</td>
</tr>
<tr>
<td>3</td>
<td>0.0968</td>
<td>0.1053</td>
<td>0.1043</td>
</tr>
<tr>
<td>4</td>
<td>0.1152</td>
<td>0.0984</td>
<td>0.1003</td>
</tr>
<tr>
<td>5</td>
<td>0.0783</td>
<td>0.0924</td>
<td>0.0967</td>
</tr>
<tr>
<td>6</td>
<td>0.0968</td>
<td>0.0872</td>
<td>0.0934</td>
</tr>
<tr>
<td>7</td>
<td>0.0737</td>
<td>0.0825</td>
<td>0.0904</td>
</tr>
<tr>
<td>8</td>
<td>0.0876</td>
<td>0.0782</td>
<td>0.0876</td>
</tr>
<tr>
<td>9</td>
<td>0.0737</td>
<td>0.0742</td>
<td>0.0850</td>
</tr>
</tbody>
</table>

In Table 1 below, the frequencies of each figure in the considered dataset, regarding the second digit of address numbers, are listed.

The BDS’s theoretical values seem slightly better; in particular the frequency range is higher, both for the observed data and for BDS’s theoretical values.

4.2 Plotting tools

The BeyondBenford package provides plotting tools to perform the comparison between the two models with the function digit.distr. In addition to arguments that are shared with dat.distr (dat, dig, upbound, main), the digit.distr function has the following additional arguments:

- **mod**: if mod="ben", the data histogram and that of Benford are displayed, if mod="ben&blo", the data histogram, that of Benford and that of BDS are plotted, and otherwise the data histogram and that of BDS are given. Default value: mod="ben".
- **col**: a vector containing two colors used to fill the bars of the histogram, if mod="ben". Default value: col=c("#FFFFAA", "#AAFFAA").
- **colbebl**: a vector containing three colors used to fill the bars of the histogram, if mod="ben&blo". Default value: colbebl=c("#FFFFAA", "#AAFFAA", "#AAFFFF").
- **colbl**: a vector containing two colors used to fill the bars of the histogram, if the latter case. Default value: colbl=c("#FFFFAA", "#AAFFFF").
- **legend**: if legend=TRUE, the legend is displayed. Default value: legend=TRUE.
- **leg**: a two-components vector containing text appearing in the legend, if mod="ben". Default value: leg=c("Observed", "Benford").
- **legbebl**: a three-components vector containing text appearing in the legend, if mod="ben&blo". Default value: legbebl=c("Observed", "Benford", "Blondeau").
- **legbl**: a two-components vector containing text appearing in the legend, if the latter case. Default value: legbl=c("Observed", "Blondeau").

Let us apply the digit.distr function to the address_PierreBuffiere dataset. The output from dat.distr is a histogram of theoretical and experimental digit distribution (Figure 3).

Example 6. digit.distr(address_PierreBuffiere, dig=2, mod="ben&blo")

Naturally, Figure 3 is consistent with Table 1.

4.3 Pearson’s chi-squared test

To quantify the quality of theoretical models, we use Pearson’s chi-squared test of goodness of fit [Pea00]: the null hypothesis states that the studied distribution is consistent with the considered theoretical distribution, i.e. Benford’s or BDS’s ones. The function chi2 determines the test statistic and its associated p-value. In addition to arguments that are shared with dat.distr (dat, dig, upbound), the chi2 function has the following specific arguments:

- **mod**: if mod='ben', the theoretical distribution considered is that of Benford, else it is BDS’s ones which is chosen. Default value: mod='ben'.

5
○ **pval**: if pval=0, the p-value is not returned, else it is available. Default value: pval=0.

Let us apply this function to the `address_PierreBuffiere` dataset. The output from `chi2` is a data frame containing the Pearson's chi-squared statistic (and the associated p-value if requested).

**Example 7.** ## Measure of Benford's Law goodness of fit

```r
chi2(address_PierreBuffiere, dig=2, pval=1)
```

<table>
<thead>
<tr>
<th>Chi2 value</th>
<th>The p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.84793221030181</td>
<td>0.921132993758269</td>
</tr>
</tbody>
</table>

## Measure of BDS's Law goodness of fit

```r
chi2(address_PierreBuffiere, dig=2, pval=1, mod="BDS")
```

<table>
<thead>
<tr>
<th>Chi2 value</th>
<th>The p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.47996278328848</td>
<td>0.98147506807756</td>
</tr>
</tbody>
</table>

In both cases the null hypothesis cannot be rejected: the studied distribution is consistent with the theoretical distributions. It can be noted that the quality of the adjustment seems slightly better with BDS's model.

5 Conclusion

The use of Benford's Law has increased rapidly in the last few years in extremely diverse fields such as mathematics, physics, biology, economics and demography, to name but a few. But the adjustment proposed by Benford is often only approximate. In some precisely described cases, it is BDS's probability distributions that are preferred. Indeed the probabilities of occurrence of digits in these distributions fluctuate around Benford's values [BDS19, BDS18].

The **BeyondBenford** package is thus a relevant tool to compare the goodness of fit of Benford’s and BDS’s distributions in a given collection of data and to offer new laws to find a better
approximation of digits distribution in the considered dataset.

References


[Pea00] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(302):157–175, 1900.


