Breakdown of Benford's law for birth data

M. Ausloos [a,b,*], C. Herteliu [c], B. Ileanu [c]

[a] e-Humanities Group, KNAW, Joan Muyskenweg 25, 1096 CJ Amsterdam, The Netherlands
[b] GRAPES, rue de la Belle Jardiniere, B-4031 Liege, Federation Wallonie-Bruxelles, Belgium
[c] The Bucharest University of Economic Studies, Bucharest, Romania

H I G H L I G H T S

- Very long survival daily time series (over 97 years) are investigated from Benford's law point of view.
- We are distinguishing between Orthodox and Non-Orthodox families in Romania.
- The results point to a drastic breakdown of Benford's law.
- Interpretation is based on the statistical aspects due to population sizes.
- This Benford's law breakdown clearly points to natural causes.

A R T I C L E   I N F O

A B S T R A C T

Long birth time series for Romania are investigated from Benford's law point of view, distinguishing between families with a religious (Orthodox and Non-Orthodox) affiliation. The data extend from Jan. 01, 1905 till Dec. 31, 2001, i.e. over 97 years or 35 429 days. The results point to a drastic breakdown of Benford's law. Some interpretation is proposed, based on the statistical aspects due to population sizes, rather than on human thought constraints when the law breakdown is usually expected. Benford's law breakdown clearly points to natural causes.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Newcomb [1] and later Benford [2] observed that the occurrence of significant digits in many data sets is *not* uniform but tends to follow a logarithmic distribution such that the smaller digits appear as the first significant digits more frequently than the larger ones, i.e.,

$$N_d = N \ \log_{10}\left(1 + \frac{1}{d}\right), \quad d = 1, 2, 3, \ldots, 9 \tag{1}$$

where N is the total number of considered 1st digits for checking the law, in short, the number of data points, and N_d is the number of the observed integer d ($= 1, 2, 3, \ldots, 9$). Usually, it seems that Benford's law breaks down when there is human manipulation or control (in various ways) of the data.

* Corresponding author at: GRAPES, rue de la Belle Jardiniere, B-4031 Liege, Federation Wallonie-Bruxelles, Belgium.
  E-mail addresses: marcel.ausloos@ulg.ac.be (M. Ausloos), claudiu.herteliu@gmail.com (C. Herteliu), ileanub@yahoo.com (B. Ileanu).

The literature on the subject is enormous [3,4] and not all papers can be quoted here. A few of socio-econo-statistical physics papers of interests are pointed out in Section 2.

In this paper, our goal is to investigate whether Benford's law holds, on long birth time series, distinguishing between the religious adhesion (Eastern Orthodox or not) of families in Romania for a time interval extending from Jan. 01, 1905 till Dec. 31, 2001, i.e. over 97 years or 35 429 days. The results point to a drastic breakdown of Benford's law. Some interpretation is proposed, based on the statistical aspects due to population sizes, rather than on human thought constraints.

In Section 3, the data acquisition is recalled. It leads to a set of time series. The data of interest are displayed through histograms and discussed following a statistical analysis in Section 3.2. All Benford law tests are found in Section 4, in particular with a test of Benford's law for the 1st and 2nd digits of the time series of the daily birth number of babies in Romania, distinguishing between Orthodox and Non-Orthodox families.

Since the results point to a drastic breakdown of Benford's law, a discussion of the findings, followed by an explanation, is found in Section 5.

Section 6 serves for a conclusion emphasizing (i) the interest of such a data study along Benford's law concepts, and (ii) the complexity of studying a community, and its religiosity, through its baby birth history.

2. Benford's law: a short state of the art literature review

The applications of Benford's law are too numerous to be all quoted here [3,4]. Nevertheless, for shining some light on the subject, we point to those Benford's law showing detection of data anomalies in actuarial and financial cases [5–10], and also in political cases [11,12] and surveys [13].

Beside these fields of applications, Benford's law has been applied in less dwelled subjects, e.g., when discussing the appearance of numbers on the internet [14], or recently, [15] for comparing articles of scientific journals. The law has been suggested to be also useful for optimizing the size of computer files [16] or for enhancing computing speed [17].

In biological sciences, Benford's law has been utilized to check the veracity of the data on clinical trials [18] and discovery of drugs [19], and in the study of diseases and genes [20]. Similarly, in physics, Benford's law has been used to detect data anomalies in numerical data on physical constants [21], atomic spectra [22,23], decay width of hadrons [24], magnitude and depth of earthquakes [25], while in astrophysics [26], for the mantissa distribution of pulsars [27] or the distances of galaxies and stars [28].

In econophysics, the yearly financial reports of the Belgian Antoinist community, income and expenses, were examined along the so-called Benford's law in order to detect any wrongdoing in the finance of such a sometimes called religious sect. Nothing anomalous was found. Note that an imperfect ("generalized") Benford's law-like form, better suited for distributions presenting a minimum at some intermediate digit, was presented in Refs. [29].

In socio-physics, closely related to our subject, some data analysis distinguishing between religious adhesion, Mir [30,31] investigated whether regularities or anomalies exist in numerical data on the country-wise adherent distribution of seven major world religions along Benford's law.

For completeness, basic (mathematical) considerations are found in Refs. [32–36].

3. Data

The data were obtained from 1992 and 2002 censuses by the Romanian National Institute of Statistics (NIS). The data of interest pertain to the record of the total number of births in Romania for persons still alive at the 1992 and 2002 census reference points. In this respect, the data might rather be called "survival occurrence from birth date". The fact that the true daily birth data are not known to us is irrelevant for the present considerations. Thereafter, we will use, for conciseness, the vocabulary "born per day" instead of the "number of births on a given day for persons still alive in 1991 and 2001, respectively, for the 1992 and 2002 census".

The data census allows to distinguish the population under various criteria, e.g. the religious adhesion. Such data must be taken with caution [30,31,37,38], but the orders of magnitude are usually trustworthy [39–41]. The most important community, from the religion point of view, according to the census sources is the Eastern-Orthodox (86.8%). The so called here below "Non-Orthodox families" (13.2%) are mainly made of Roman-Catholics (4.7%), Reformed Church (3.2%), Pentecostals (1.5%), Greek-Catholics (0.9%), Baptists (0.6%), Seventh Day Adventists (0.4%), Muslims (0.3%), Unitarians (0.3%), Lutherans (0.3%), Evangelicals (0.2%) and Old Rite Christians (0.2%). Other denominations (0.6%), including atheists, have each a smaller size.

3.1. Time series

Thus, we have some information on births in Romania from Jan. 01, 1905 till Dec. 31, 2001, i.e. 97 years or 35 429 days, see Fig. 1. Note that several data points overlap in the display, because the number of born babies occurs to be the same on various days during the year. It is also observed that several occurrences seem to exist as outliers. We have kept the official data unchanged.