sciendo Journal of Social and Economic Statistics

Vol. 8, No. 2, Winter 2019
DOI: 10.2478/jses-2019-0001

DETECTING ANOMALOUS DATA IN HOUSEHOLD SURVEYS: EVIDENCE FOR ARGENTINA

Fernando Antonio Ignacio González[a]

Abstract

This paper advances in the detection of anomalous data in income reports of Argentina. In particular, income declared by households surveyed in the Encuesta Permanente de Hogares (EPH, Permanent Household Survey in English) -for the period 2003-2017- and in the Encuesta Anual de Hogares Urbanos (EAHU, Annual Urban Household Survey in English) -for the period 2010-2014- are analyzed.

A widely known technique in forensic accounting and auditing, such as Benford's law -also known as the first digit law- is used. If the analyzed data were generated naturally -free of manipulation- it should follow the logarithmic distribution of Benford. The Chi-square test and the absolute mean deviation (MAD) are used for verification.

The results suggest that the income reported in the EPH does not follow the Benford distribution and the degree of compliance with this law decreases significantly between 2007-2015 coinciding with the intervention period of the Instituto Nacional de Estadísticas y Censos (INDEC, National Institute of Statistics and Censuses in English).

Keywords: Income, Household surveys, Benford´s law.
JEL: M42, M48

Author's Affiliation

[a] Instituto de Investigaciones Económicas y Sociales del Sur, UNS-CONICET. E-mail: fernando_gonzalez01@hotmail.com

1. Introduction

Household survey microdata are widely used by social scientists -when available- and, in particular, information referring to household income is of special relevance in topics such as poverty or inequality. Thus, the availability of quality information, tamper free, is vital as support for research.

In recent decades, household surveys, carried out by national statistical institutes, have proliferated in Latin America as means of data collection. In particular, Argentina has an extensive tradition in conducting household surveys: the EPH has been carried out since 1973 and, at present, is carried out on a quarterly basis covering 31 urban agglomerates. Also, the EAHU was conducted between 2010-2014 with a wider geographical coverage than the EPH.

Information referring to income -personal or from household- has received special attention, not only because of its importance on estimations, but also due to the presence of sampling problems. On the one hand, it is frequently the category with the most missing values in household surveys and, in addition, what has been declared may be under-estimated -especially in higher income households- (Medeiros, Castro *et al*., 2018). Thus, INDEC has explored different alternatives to address the above: from 2003 to 2015 an imputation of income was made -through the random hot deck method- for those who had not answered the income questions. Since 2016, INDEC opted to correct with reweighting for non-response instead of income imputation.

In the Argentine case, the lack of credibility of publications made by INDEC between 2007-2015 is publicly known. Indeed, between 2007 and 2015, INDEC was intervened and its reports for this period should be considered with caution (INDEC, 2016). The intervention consisted of the displacement of career officials (Minoldo and Born, 2019) and which was followed by problems such as omission of geographical coverage, discrepancy in population projections, lack of operational training of personnel, biased practices, among others (INDEC, 2016). The Economic Commission for Latin America and the Caribbean (ECLAC) recognizes that, in this period, a sub-estimation of inflation took place, which led to a reduction in poverty estimates (ECLAC, 2018).

Different statistical tools can be used to examine the existence of anomalous situations in income statements. Benford's law is a widely disseminated tool used to detect whether a data distribution behaves naturally or not. Indeed, Benford's law analyzes the frequency of occurrence of the first significant digits (FSD) of a distribution of interest, which are compared to a theoretical distribution. In a completely random distribution the occurrence of each significant first digit would be the same but, interestingly, Benford's law suggests that the number 1 as the first digit is more frequent than the 2 and thus, the probabilities of occurrence of each first digit are decreasing up to 9. This tool has been applied to the detection of tax fraud (Nigrini, 1996), manipulation of GDP macroeconomic data (Holz, 2014), household income (Finn and Ranchhod, 2013; Villas- Boas, Quizi *et al*., 2015; 2017), public spending (Cella and Zanolla, 2018), among others.

Considering the above, this work advances in the analysis of self-reported household income data for Argentina examining whether its distribution occurs naturally -in compliance with Benford's Law- or not. To do this, Section 2 formally defines Benford's law and examines relevant background for the analysed case. It also presents sources of information and formal tests used. Section 3 discusses main results. Finally, section 4 describes conclusions.

2. Materials and methods

2.1 Benford Law and its applications

Benford's law, also known as first digit or Newcomb-Benford law, was initially reported by Newcomb (1881) who noted that the first pages of certain books tended to be more worn than later pages. Years later, Benford (1938) re-examined this regularity and verified its compliance in distributions such as physical constants, household addresses, river lengths, etc.

Benford suggested that the occurrence of the first significant digits follows a logarithmic distribution:

$$P_d = \log_{10}(1 + \frac{1}{d}) \; con \; d = 1, 2, \dots, 9 \qquad (1)$$

where P_d is the probability for a given number of having *d* as the first significant digit. Thus, the probability of occurrence of each digit is as follows:

Table 1: Occurrence of each significant digit according to Benford's law

Digit	Probability of occurrence
1	0.3010
2	0.1761
3	0.1249
4	0.0969
5	0.0792
6	0.0669
7	0.0580
8	0.0511
9	0.0458

Source: own elaboration

This law presents the interesting property of scale invariance (Mir, Ausloos *et al.*, 2014), that is, changes in the unit of measurement of the data of interest do not affect compliance with this law.

The intuition behind this law is that -given the empirical verification that lower digits tend to occur more frequently than higher digits[1]- if a set of data was generated naturally it will tend to comply with Benford's theoretical distribution. Failure to comply with this law suggests the possible existence of data manipulation and which should be verified with a detailed analysis of the information in each case. Therefore, this tool has been widely used in areas such as auditing, forensic accounting and, in general, applied to data distributions related to human behavior.

Logically, not all data distributions comply with what is suggested by Benford's law. In this regard, those data sets subject to truncation, censorship or rounding tend to breach it. Those coded distributions -e.g.: telephone numbers or postal codes- also tend to default (Gunnel and Todter, 2007).

Several works have tried to verify compliance with this law in the context of self-reported income. Nigrini (1996) inquires about the existence of tax fraud in tax returns in the United States and finds that small taxpayers tend to manipulate their returns more frequently than large taxpayers. Referred to financial reports of US companies in 2012, Henselmann, Scherr *et al.*

[1] Benford's law was observed empirically by Newcomb (1881) and Benford (1938) but its statistical formalization was presented in Hill (1995) using the central limit theorem.

(2012) analyze compliance with Benford's law and find that, in general, the distributions fit the theoretical distribution well -with exceptions such as the Income category, subject to possible fraud-.

In the particular case of household surveys, Judge and Schechter (2009) analyze compliance with the law in 9 surveys of developed and developing countries related to the amount of crops harvested and animals per household. They find that surveys of developed countries (United States) tend to approximate Benford's law better, compared to surveys of developing countries (Paraguay, Mexico, Pakistan, etc.) and attribute it to the fact that in the first case respondents answer more accurately since they consult written records at the time of the survey. Referred to income declared in household surveys, Villas-Boas, Quizi *et al.* (2017) analyze compliance with Benford's law considering microdata from 13 countries in Europe and find that in no case is it possible to reject compliance with this law.

Benford's law has not been without criticism. On the one hand, its non-compliance for a certain distribution does not imply *per se* the existence of fraud or manipulation, but invites a more detailed analysis of the data. In addition, not all distributions tend to follow Benford´s logarithmic distribution, even in the absence of manipulation -for some of the reasons mentioned above (censorship, truncation, coding, small samples, etc.)-. In any case, Benford's law is an intuitive and easy-to-implement tool -especially in large volumes of data- as the first approach in the detection of anomalous data. Also, it has been suggested that, although a distribution does not comply with Benford's theoretical distribution, its degree of concordance over time can be analyzed (Nigrini, 2017): that is, a distribution can always break this law, but it would be anomalous for periods where non-compliance occurs to a greater extent.

2.2 Formal tests

Logically, compliance or not with Benford's law must be corroborated based on some formal proof. Indeed, Pearson's Chi-square test (Pearson, 1900) is widely used to test whether a sample comes from a distribution with a certain probability density function. The Chi-square statistic is calculated as:

$$Chi - cuadrado = \sum_{i=1}^{N} \frac{(f_i - e_i)^2}{e_i} \qquad (2)$$

where N is the number of existing classes, the subscript i denotes each of the classes, f_i is the observed frequency and e_i is the expected frequency. For arbitrarily large values of the statistician, the null hypothesis will tend to be rejected and therefore, it is concluded that the analysed distribution does not come from a Benford´s distribution. The Chi-square test can present problems of type I error in large samples and therefore reject compliance with Benford's law in situations where the difference in proportions is reduced (Druica, Oansea *et al.*, 2018). The same goes for Z or Kolmogorov-Smirnov type tests (Barney and Schulzke, 2016).

Trying to overcome the above, a measure used to evaluate the degree of concordance of a distribution with Benford's law is the mean absolute deviation (MAD) defined as (Barney and Schulzke, 2016):

$$MAD = \frac{1}{N}\sum_{i=1}^{N} |p_i(o) - p_i(e)| \quad (3)$$

where $p_i(o)$ is the proportion of observations observed for class i, $p_i(e)$ is the expected proportion for class i according to Benford's law. Close to zero values of the MAD suggest high compliance with Benford's law. In this regard, Drake and Nigrini (2000) suggest a critical threshold of 0,012 and which is followed in this work[2].

2.3 Sources of information

First of all, two different household surveys are considered. The EPH is considered for the 2003-2017 period while, the EAHU is considered for the 2010-2014 period. Both surveys are published by INDEC (INDEC, 2019). Income variables of the main occupation and other occupations are examined

3. Results

The results for income reported in the EPH, for selected years, are presented below:

Table 2: Benford´s law conformity, Chi² and MAD, in EPH

Year	Income	Chi-square	p-value	MAD
2003	Income main occupation	1058.27	0.000	0.017
	Income others occupations	223.61	0.000	0.017
2006	Income main occupation	4587.59	0.000	0.0204
	Income others occupations	291.99	0.000	0.015
2009	Income main occupation	6033.69	0.000	0.0253
	Income others occupations	380.37	0.000	0.0163
2012	Income main occupation	10107.23	0.000	0.0329
	Income others occupations	388.19	0.000	0.0187
2015	Income main occupation	4879.23	0.000	0.028
	Income others occupations	186.99	0.000	0.0176
2017	Income main occupation	5857.29	0.000	0.018
	Income others occupations	145.37	0.000	0.0096

Source: own elaboration based on EPH.

It is observed that, in no year the reports of personal income of the EPH follow the distribution of Benford according to the Chi-square statistic. When using the MAD, at least striking results are obtained. First, the distribution of income from *other occupations* presents a greater degree of conformity with Benford's law than the distribution of income from the

[2] Drake and Nigrini (2000) propose a conformity scale: between 0-0.004 close conformity; between 0.004-0.008 acceptable conformity; between 0.008-0.012 marginally acceptable conformity; greater than 0.012 disconformity.

main occupation. Second, in the 2006-2015[3] period, the MAD increases significantly for both distributions and moves away from the critical threshold of 0.012 of Drake and Nigrini (2000). Assuming that the possible distortions introduced by the respondents remained stable during the period, the above suggests that the intervention of INDEC reached the income collected in the EPH -i.e. income data was tampered-. This is shown in the following graph:

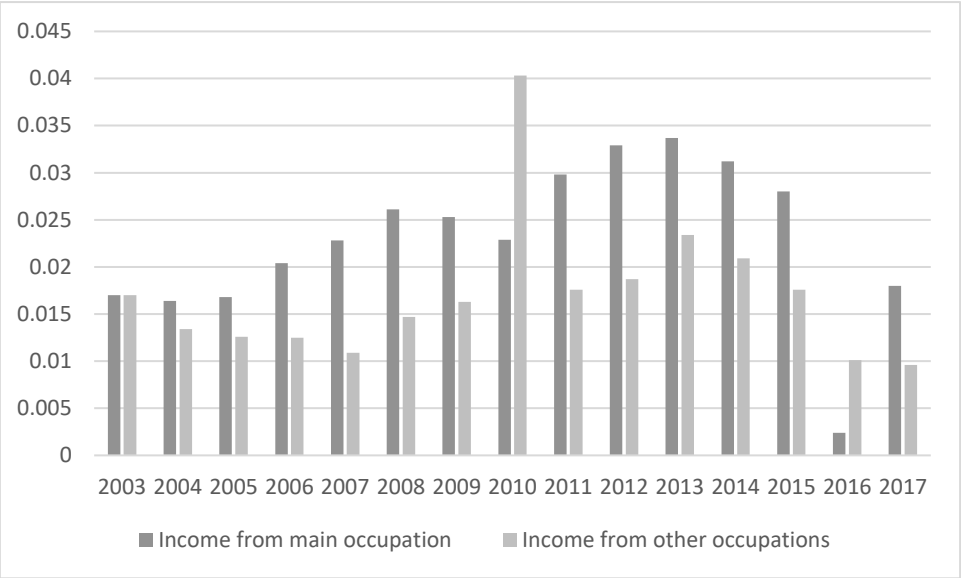
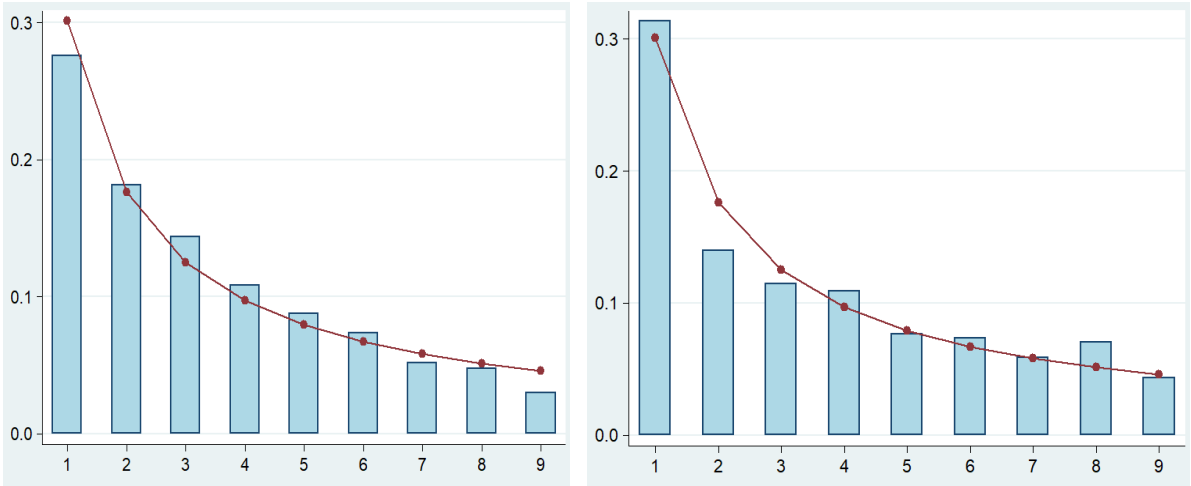


Figure 1: MAD for EPH personal income, 2003-2017
Source: own elaboration based on EPH-INDEC.

Logically, non-compliance with Benford's law implies that the occurrence of each first digit represents a different proportion to that predicted by the theoretical distribution. In this regard, distributions are plotted year by year.



[3] Each quarterly wave of the EPH is published with a lag of four months. Given that the INDEC intervention took place since January 2007, it is expected that if the MAD increase was a consequence of this intervention, its effects can be seen since 2006 results.

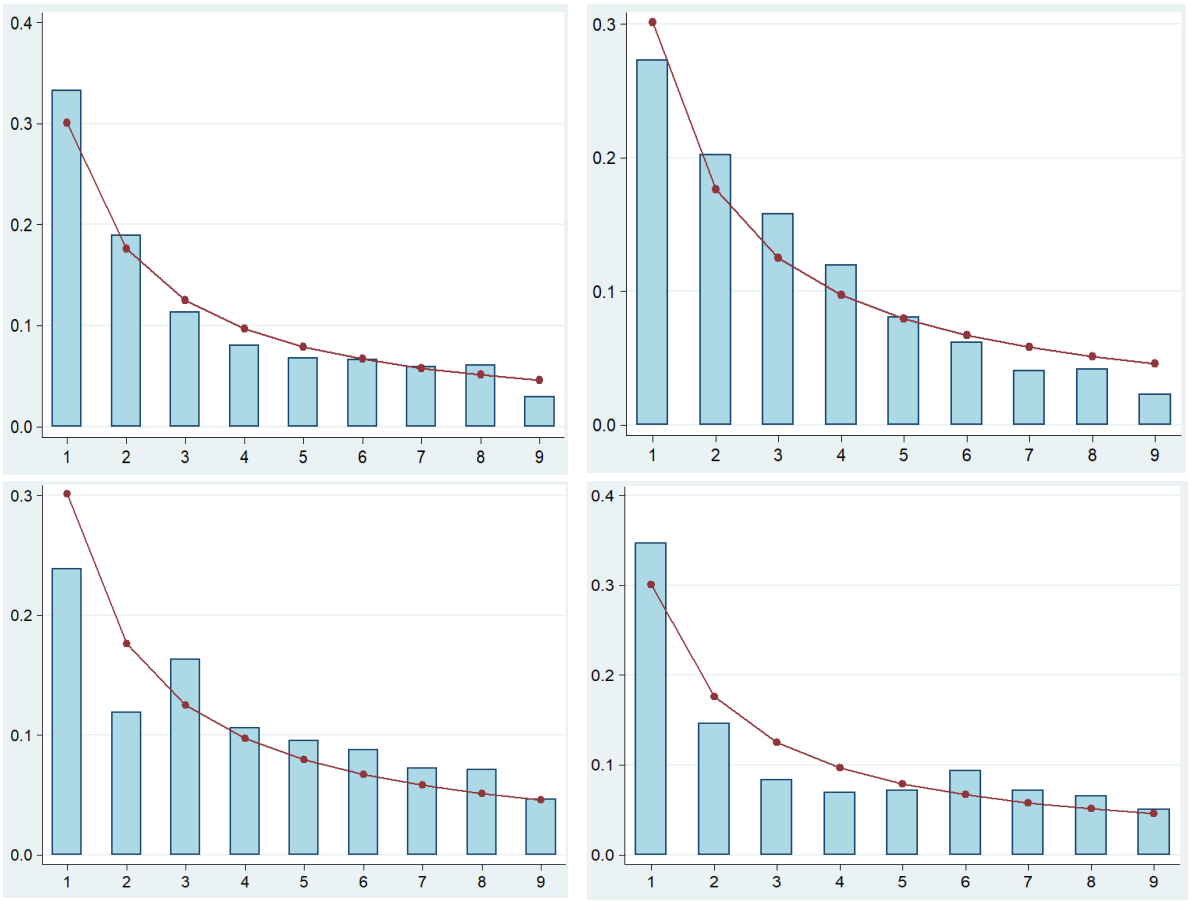


Figure 2: Probability of occurrence of each first digit in EPH and Benford's law

Source: own elaboration based on EPH-INDEC.

Note: the first row corresponds to years 2003 (left) and 2006 (right), the second row to years 2009 and 2012 and, the third row to years 2015 and 2017.

These figures do not show a pattern of under/over-representation in the occurrence of the first digits. Thus, the digit 1 appears under-represented in the years 2003, 2012 and 2015, while the opposite occurs in the remaining years.

Considering results from EAHU, it is possible to compare compliance with Benford's law between those cases where income imputation was made and those where it was not performed.

Table 3: Compliance with Benford's law and income imputation

			Chi-square	p-value
2010	Income from main ocupation	Imputed	689.53	0.0000
		Non imputed	50644.32	0.0000
	Income from other ocupations	Imputed	6922.99	0.0000
		Non imputed	104987.68	0.0000
2012	Income from main ocupation	Imputed	1040.30	0.0000
		Non imputed	46510.87	0.0000
	Income from other ocupations	Imputed	7995.74	0.0000
		Non imputed	98165.24	0.0000
2014	Income from main ocupation	Imputed	737.06	0.0000
		Non imputed	39774.82	0.0000
	Income from other ocupations	Imputed	6821.49	0.0000
		Non imputed	84037.90	0.0000

Source: own elaboration based on EAHU.

From the table above, the lack of compliance with Benford's law can be observed in all the analysed cases. Thus seen, income imputation does not seem to be the source of the discrepancies. In addition, there are no differences -according to EAHU data- between income from the main occupation and those from other occupations.

4. Conclusions

Throughout this paper, attempts have been made to verify whether self-reported income data, in Argentina, are presented naturally -or not- according to what is proposed by Benford's law -also known as the first-digit law-. If verified, the distributions of interest should follow a logarithmic distribution in which minor digits are more likely to occur than the larger digits.

To verify the above, the use of the Chi-square test -which is a high power test- was widely observed. Also, it may be problematic for large samples (Barney and Schulzke, 2016; Druica, Oansea *et al*., 2018) and is usually complemented by measures such as absolute mean deviation (MAD).

In the case of income reported in household surveys, through the EPH, reduced compliance with Benford's law was observed. The Chi-square test rejects, in all years, the null hypothesis of conformity. An exception is the distribution of income from *other occupations* for those in a condition of employer. In terms of the MAD, values close to the critical threshold were observed in 2004-2007 and 2016-2017 -especially for the distribution of *other occupations*-. Interestingly, the values of the MAD increase and move widely away from the threshold of Drake and Nigrini (2000) during the period of the intervention of INDEC (2007-2015), and then decrease in the stage of re-normalization of the organism. The above is a first statistical evidence of the -widely disseminated- idea of manipulation in income statistics in that period. However, further analysis is necessary to rule out other explanations of the observed.

The EAHU data were analyzed to test for differences between imputed and non-imputed personal income. The results suggest the absence of significant differences between both cases.

In the future, it is relevant to continue examining the income reported in the EPH with a greater time horizon to corroborate, or discard, the manipulation hypothesis during the period of the intervention of INDEC.

References

Barney, B. and Schulzke, K. (2016). Moderating "cry wolf" events with excess MAD on Benford´s law research and practice. *Journal of forensic accounting research*, 1(1), pp. 66-90.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American philosophical society*, 78(4), pp. 551-572.

Cella, R. and Zanolla, E. (2018). Benford´s law and transparency: an analysis of municipal expenditure. *Brazilian business review*, 15(4), pp. 331-347.

ECLAC (2018). *Medición de la pobreza por ingresos. Actualización metodológica y resultados*. Metodologías de la CEPAL N° 2. Available in: https://repositorio.cepal.org/bitstream/handle/11362/44314/1/S1800852_es.pdf

Drake, P. and Nigrini, M. (2000). Computer assisted analytical procedures using Benford´s law. *Journal of accounting education*, 18, pp. 127-146.

Druica, E., Oancea, B. and Välsan, C. (2018). Benford´s law and the limits of digit analysis. *International Journal of Accounting Information Systems*, 31, pp. 75-82.

Finn, A. and Ranchhod, V. (2013). *Genuine fakes: the prevalence and implications of fieldworker fraud in a large South African survey*. Working paper series 115, Southern Africa labour and development research unit.

Gunnel, S. and Todter, K. (2007). *Does Benford´s law hold in economic research and forescasting?*. Discussion paper 32, Deutsche Bundesbank.

Henselmann, K., Scherr, E. and Ditter, D. (2012). *Applying Benford´s law to individual financial reports: an empirical investigation on the basis of SEC XBRL filings.* Working papers in accounting valuation auditing 2012-1 [rev.], FriedrichAlexander-Universität Erlangen-Nürnberg, Lehrstuhl für Rechnungswesen und Prüfungswesen, Nürnberg.

Hill, T. (1995). A statistical derivation of the significant-digit law. *Statistical Science*, 10(4), pp. 354-363.

Holz, C. (2014). The quality of China´s GDP statistics. *China Economic Review*, 30, pp. 309-338.

INDEC (2016). *Mercado de trabajo: principales indicadores. Segundo trimestre de 2016. Consideraciones sobre la revisión, evaluación y recuperación de la Encuesta Permanente de Hogares (EPH). Anexo Informe de Prensa*. Available in: https://www.indec.gob.ar/ftp/cuadros/sociedad/anexo_informe_eph_23_08_16.pdf

INDEC (2019). Bases de datos del mercado laboral: *Encuesta Permanente de Hogares y Encuesta Anual de Hogares Urbanos.* Available in: https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos

Judge, G. and Schechter, L. (2009). Detecting problems in survey data using benford´s law. *Journal of human resources*, 44, pp. 1-24.

Medeiros, M., Castro, J. and Azevedo, L. (2018). Correcting the underestimation of top incomes: combining data from income tax reports and the Brazilian 2010 census. *Social indicators research*, 135(1), pp. 233-244.

Minoldo, S. and Born, D. (2019). *Claroscuros. 9 años de datos bajo sospecha*. Buenos Aires: ESEditora.

Mir, T., Ausloos, M. and Cerquetti, R. (2014). Benford´ law predicted digit distribution of aggregated income taxes: the surprising conformity of Italian cities and regions. *Physics and society,* 87, pp. 261-279.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American journal of mathematics*, 4(1), pp. 39-40.

Nigrini, M. (1996). A taxpayer compliance application of Benford´s law. *Journal of the American Taxation association*, 18, pp. 72-92.

Nigrini, M. (2017). Audit sampling using Benford´s law: A review of the literature with some new perspectives. *Journal of emerging technologies in accounting*, 14(2), pp. 29-46.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables in such that it can be reasonably supposed to have arisen form random sampling. *The London, Edinburgh and Dublin Philosophical magazine and journal of science,* 50(302), pp. 157-175.

Villas-Boas, S., Quizi, F. and Judge, G. (2015). Is Benford´s law a universal behavioral theory? *Econometrics,* 3, pp. 698-708.

Villas-Boas, S., Quizi, F. and Judge, G. (2017). Benford´s Law and the FSD distribution of economic behavioral micro data. *Physisca A*, 486, pp. 711-719.