Benford's Law as an Indicator of Survey Reliability
Can We Trust Our Data?

Kaiser, Micha

Document Version
Final published version

Published in:
Journal of Economic Surveys

DOI:
10.1111/joes.12338

Publication date:
2019

JOURNAL OF ECONOMIC SURVEYS

doi: 10.1111/joes.12338

BENFORD'S LAW AS AN INDICATOR OF SURVEY RELIABILITY—CAN WE TRUST OUR DATA?

Micha Kaiser*

University of Hohenheim
and Center for Consumer, Markets, and Politics (CCMP)

Abstract. This paper analyzes how closely different income measures conform to Benford's law, a mathematical predictor of probable first digit distribution across many sets of numbers. Because Benford's law can be used to test data set reliability, we use a Benford analysis to assess the quality of six widely used survey data sets. Our findings indicate that although income generally obeys Benford's law, almost all the data sets show substantial discrepancies from it, which we interpret as a strong indicator of reliability issues in the survey data. This result is confirmed by a simulation, which demonstrates that household level income data do not manifest the same poor performance as individual level data. This finding implies that researchers should focus on household level characteristics whenever possible to reduce observation errors.

Keywords. Benford´s law; Data quality; Fraud detection; Measurement error; Survey quality

JEL Classification. C18; C15; C46; C55; C81; I100

1. Introduction

The widespread use of survey data across both social and life sciences has led to the development over recent decades of a multitude of econometric and statistical methods designed to detect causal relations and make the most precise predictions possible. One crucial issue that remains, however, is measurement error, which may severely bias estimates.[1] Although research streams in both econometrics and statistics are already concentrating on how to deal with measurement error-induced reliability problems, researchers could benefit greatly from advance knowledge about data quality.[2] Such knowledge is even more important for policy makers, who do not apply econometric techniques to the survey data on which they base their policy inferences, especially when those data have been collected for particular government purposes.

Although myriad methods already exist for checking the quality of a particular data set, most rely on expensive procedures like matching employee and employer data to enable comparison (for a detailed description, see Mellow and Sider, 1983; Duncan and Hill, 1985). An alternative, and less costly, option is to apply Benford's law of likely first digit distribution in a data set, and detect anomalies and possible quality problems by measuring the number of deviations from the theoretical pattern. This rule, however commonly used by tax authorities to detect fraud, is still not widely applied in social sciences.[3] Authors who have used it to evaluate (survey) data quality include Judge and Schechter (2009), mainly for agricultural data; Nigrini and Miller (2007), for hydrological data; Sandron (2002), for population numbers; Mir (2014), for religious data; Ausloos *et al.* (2015), for long-term birth numbers; Fu *et al.* (2007), for image forensics; and Swanson, Cho and Eltinge (2003), for consumption expenditures. Additionally, further studies exist that discuss the use of Benford´s law as a lie detector (Gauvrit *et al.*, 2017), to detect incorrect

*Corresponding author contact email: micha.kaiser@uni-hohenheim.de

Journal of Economic Surveys (2019) Vol. 33, No. 5, pp. 1602–1618

information of countries regarding their effort to combat money laundering (Deleanu, 2017), or to assess the reliability of financial reports in developing countries (Shi *et al.*, 2018). The literature thus lacks large-scale analyses of how individual income data conforms to Benford's law, knowledge that could be used to improve the assessment of survey data quality. Given the widespread use of income data in both econometric and policy analyses, however, correct assessment of data quality is crucial for accurate inference.

This paper attempts to meet this need by making the following contributions to the existing literature: First, we show plausible reasons to consider income as generally in compliance with Benford's law. Second, we use the law to assess income data quality in six harmonized survey data sets that are widely used in economics and social science. Third, by using three different variables, we detect systematic differences in the quality and design of these income measures. Fourth, we introduce a simple but efficient simulation algorithm that improves the validity of a Benford analysis for any particular survey data set.

The paper is organized as follows: Section 2 describes Benford's law and the conditions for its applications, explains the analytical method used for our analysis, and describes the data sets analyzed. Section 3 reports the main results and Section 4 presents the specifications for and outcomes of the simulation test for robustness. Section 5 then discusses the results and concludes the paper.

2. Methods and Data

2.1 *Benford's Law*

Although named for American physicist Frank Benford (1938), the phenomenon on which Benford's law is based was first reported by Francis Newcomb (1881), who noted a more frequent use of logarithmic tables that included numbers beginning with low digits. From this observation, he derived a mathematical rule for the probability p of first digits d occurring in the numbers of a given data set. This rule is characterized by the following logarithmic function (with B as the logarithmic base), which empirically predicts the occurrence of first digits in a broad variety of data sets:

$$p\ (j) = log_B\ \left(1 + \frac{1}{j}\right) \tag{1}$$

Benford (1938) independently made this same observation over a half century later and published his own first-digit law.[4]

In Figure 1, we illustrate Benford's law by mapping first digits $j \in \{1, 2, 3 \ldots, 9\}$ onto a probability space to produces a monotonically decreasing graph with higher probability values for lower digits and almost uniform values for higher digits. Whereas the probability of observing a number beginning with $j = 1$ in a given data set is approximately 6.5 times higher than that for a number beginning with $j = 9$, the probability for a number beginning with $j = 8$ (rather than $j = 9$) is only 1.1 times higher.

Benford's law does not, however, characterize every data set: rather, its occurrence requires the presence of various criteria. Pinkham (1961), for instance, shows that changing the measurement scale should not change the first digit distribution in a set of numbers. Thus, if the first digit occurrence probability in a data set expressed in kilometers changes when the same data are expressed in meters, the data are probably not following Benford's law. In subsequent work, Hill (1995a) further shows that Benford's law is characterized by base invariance, meaning that the first digit occurrence probability must not change with a change in the base B of the underlying logarithmic function (see equation (1)). Moreover, as Pietronero *et al.* (2001) pointed out, the occurrence of Benford's law is a result of multiplicative processes, implying necessary Benford compliance by data generating processes that follow a Markov chain (Berger *et al.*, 2011). This latter throws valuable light on why many observable (economic) data obey Benford's law: many economics processes (e.g., GDP growth, employment rates, or income development) are describable by Markov chains (Le Gallo, 2004). Another possible explanation for this interrelation is

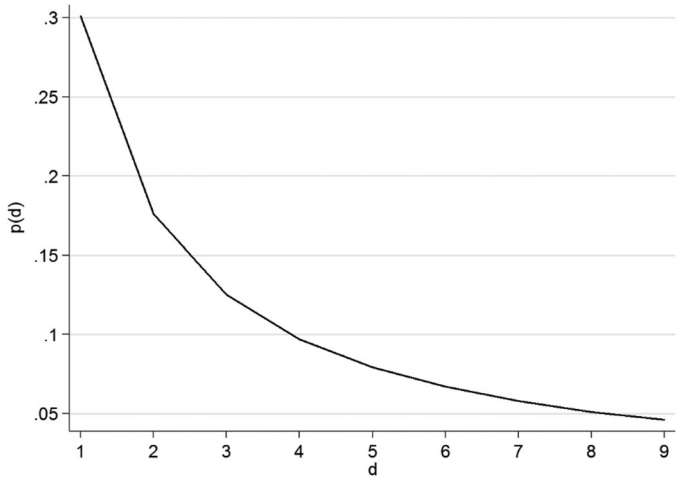

Figure 1. Benford First Digit Distribution of Numbers in a Given Data Set.

given by the fact that quantities that are related to the exponential function (as most economic processes) show necessarily the pattern described by Benford's law (Whyman *et al.*, 2016). These findings are in line with recent work of Villas-Boas *et al.* (2017), which demonstrate that not only physical but also economic behavioral systems are consistent with Benford's law.

The fact that the random numbers used in data generating processes need not be identically distributed may also explain the frequent adherence to Benford's law of "real world" data sets (Hill, 1995). For example, although the numbers used in scientific publications are invariably taken from a mix of different probability distributions, they tend to obey Benford's law to some extent (Tödter, 2009). The latter observation is also in line with Formann's (2010) simulation evidence that data from right-tailed distributions, particularly, are likely to obey Benford's law. Again, this adherence is likely if the resulting random variables in the data set stem from a mix such as the ratio of two half-normal distributions. The latter conversely implies, however, that data resulting from a symmetric distribution is unlikely to obey Benford's law, which in fact is seldom the case in economic data (e.g., income) because of its tendency to be log normally distributed (Clementi and Gallegati, 2005).

Wallace (2002) thus proposes a rule of thumb for data set adherence to Benford's law that expects fulfillment of two criteria: the mean of the data set should be higher than the median, and the data set should be characterized by a positive skewness value. This rule implies that to obey Benford's law, the data must have sufficient volatility, including a sufficiently broad range of numbers (Durtschi *et al.*, 2004). Durtschi *et al.* (2004) conversely propose the following exclusion restriction: a data set is not likely to obey Benford's Law if it is, for instance, "influenced by human thought," has a "built in minimum or maximum," or "is composed of assigned numbers" (p. 24). As a result of the above, economic data are widely accepted to generally conform to Benford's law. Hence, any discrepancies in the data sets evaluated here may indicate a serious data reliability problem.

2.2 *Hypothesis Testing*

To make the most accurate evaluations of data reliability, we incorporate several of the many methods for testing adherence to the first digit law but interpret the separate results as an aggregate. In particular, we

combine a graphic analysis with three statistical methods: Pearson's (1900) chi-squared (χ^2) test as well as the Kolmogorov–Smirnov (1948) test for goodness of fit, and a distance measure developed by Leemis *et al.* (2000).

The Pearson's goodness-of-fit uses the following test statistic:

$$\chi^2 = \sum_{j=1}^{m} \frac{\left(N_j - n_j\right)^2}{n_j}\,, \tag{2}$$

which computes the sum of the squared (and standardized) deviations of the empirical number of observations N_j (with $n = \sum_{j=1}^{9} N_j$) for each of the first digits $j \in \{1, 2, 3 \ldots, 9\}$ and the expected frequency $n_j = p(j) \times n$ as proposed by Benford's law. Given a particular significance level α, we will not reject the null hypothesis that the data obey Benford's law if the test statistic does not exceed the corresponding critical value (20.09 for $\alpha = 0.01$ and 15.51 (13.36) for $\alpha = 0.05$ ($\alpha = 0.10$), respectively).

The Kolmogorov–Smirnov (1948) test is expressed as

$$D_n = sup_x \left|F_n(x) - F(x)\right|, \tag{3}$$

where $F(x) = p(x \leq j)$ represents the cumulative distribution function of the Benford distribution and $F_n(x) = \Pr_n(x \leq j)$ denotes the empirical cumulative distribution function (cumulative frequency) for all n observations.[5] Here, we reject the null hypothesis if the test statistic D_n exceeds the critical values k, calculated as follows (Sachs, 2004, p. 427–431):

$$k = \frac{b_\alpha}{\sqrt{n}}\,, \tag{4}$$

where $b_\alpha \in \{1.224,\ 1.358,\ 1.628\}$ depends on significance level. The distance measure (Leemis *et al.*, 2000) used to test the degree of similarity between the first digit distribution in the data sets analyzed and those in the Benford distribution is then expressed by

$$m = max_{j=1,2,\ldots,9} \left\{\left|\Pr_n(j) - p(j)\right|\right\} \tag{5}$$

2.3 *Data*

The analysis evaluates six different longitudinal data sets from the Health and Retirement Study (HRS) family of studies, originated by the U.S. National Institute on Aging (NIA). All six focus on a broad range of health, wealth, and income issues, and include quality of life measures that provide insights into the life situations of older citizens. To accurately identify and compare the quality differences in the individual data sets, in five cases, instead of the originals, we employ the harmonized data sets provided by Gateway to Global Aging (2017).[6] A major advantage of harmonization is that the resulting data sets tend to include similar variables, which facilitates both the analysis and interpretation of the separate analytical results:

HRS (America). The original Health and Retirement Study (HRS), funded by the NIA, whose first wave (1992–1993) served as a baseline for the remaining 11 waves (ending in 2014–2015), which all closely mirrored its structure. From an original sample size of 12,600 individuals over 51 years old, the sample size increased to 18,700 by 2014.

Harmonized ELSA (England). The English Longitudinal Study of Ageing (ELSA), administered to individuals over 50 years old, was funded by the NIA and three different UK government departments.[7]

Its sample declined from 12,000 in 2002–2003 to approximately 9,600 in 2014–2015. The harmonized dataset covers four of the original six waves.

Harmonized SHARE (Europe).[8] The Survey of Health, Ageing, and Retirement in Europe (SHARE), structured like the HRS and ELSA and funded by the European Commission, was administered to individuals over 50 years old. Having been conducted in 19 European countries plus Israel, SHARE offers notably larger samples: 30,700 in the first wave and 68,200 in the latest. The harmonized data set covers four SHARE waves: 2004–2005, 2006–2007, 2010–2011, and 2012–2013.

Harmonized CRELES (Costa Rica). The Costa Rican Longevity and Healthy Aging Study (CRELES) was a joint project of the University of Costa Rica's Centroamericano de Población, Instituto de Investigaciones, and the University of California, Berkeley. In contrast to its sister studies, the CRELES, administered in five waves from 2004–2005 to 2012–2013, includes two different cohorts: those over 55- and those over 60 years old.

Harmonized TILDA (Ireland). The Irish Longitudinal Study on Ageing (TILDA), conducted with individuals over 50 years old, was funded by the Department of Health, Atlantic Philanthropies, and Irish Life. The harmonized TILDA data set compiles two survey waves: 2010–2011 (sample = 8,500) and 2012–2013 (sample = 7,200).

Harmonized LASI (India). The Longitudinal Aging Study in India (LASI), funded by the NIA, the Government of India, and the United Nations Population Fund, differs from the other surveys in that it was only administered once, in 2010, to individuals over 45 years old. Nonetheless, although the sample only includes 1,600 individuals, the survey is structured similarly to the HRS.

To test whether the numbers in the data sets come close to following Benford's law, we focus on three different (income) variables measured over the previous 12 months: total household income (HITOT) from all sources; total respondents earnings (RIEARN) from both labor and trade; and spousal employment earnings (SIEARN) from both labor and trade.[9] One reason for choosing these particular variables is that not all the individual surveys covered by the harmonized data sets necessarily address the same topics, which makes it hard to compare data set quality. For instance, although it would be interesting to evaluate the reliability of individual health data (e.g., hospital stays per year), this information is only available in some surveys. Moreover, the health related data contained in almost every survey tend to refer to different time spans.[10]

At the same time, because the conditions for a set of numbers (or variables) to obey Benford's law are relatively strict, not every variable can be exploited for data reliability assessment using a Benford analysis. Information on cigarette intake per day, for example, although an interesting candidate for reliability testing, does not obey Benford's law because of its built-in maximum (Durtschi *et al.*, 2004). Income, however, is a widely used analytic variable across scientific disciplines—especially in economics or social sciences—so the reliability of these income variables is central to assessing the validity of the corresponding analytic conclusions. In particular, because wealth related policy decisions tend to be based on major survey data, the data underlying income distribution information must be reliable.[11]

To avoid analytic distortion, our summaries of the mean, median, skewness, and number of observations for these three variables in each data set (Tables 1–3) exclude observations in which each variable has a zero value. For all data sets, the mean values for HITOT, RIEARN, and SIEARN (Tables 1–3, respectively) are higher than the medians, and all distributions appear positively skewed. This pattern is a strong indicator that the income variables used should generally obey Benford's law and hence be suitable for detecting reliability problems in the data (Wallace, 2002).

Table 1. Descriptive Statistics for Total Household Income for the Different Data Sets.

HITOT	Mean	Median	Skewness	Obs.
HRS	57,510.78	34,800	189.72	224,287
ELSA	23,854.1	18,616	11.463	61,742
CRELES	16,221.65	1,440	102.56	10,703
LASI	124,161.7	58,900	5.62	1,504
SHARE	36,651.25	22,933.14	5.97	25,028
TILDA	59,086.01	32,028	19.534	12,579

Notes: The values are measured in the following units: HRS – nominal dollars; ELSA – nominal pounds; CRELES – 1,000 Costa Rican colons; LASI – Indian rupees; TILDA – euros; SHARE – euros, except for Denmark, Sweden, Switzerland, Poland, Czech Republic, Hungary, and Estonia. The first wave of SHARE includes before-tax income, whereas all subsequent waves consider after-tax income.

Table 2. Descriptive Statistics for Respondent Employment Earnings for the Different Data Sets.

RIEARN	Mean	Median	Skewness	Obs.
HRS	36,446.42	25,000	30.89	84,375
ELSA	14,736.67	12,009.24	12.70	19,796
CRELES	4,015.614	1,800	6.39	2,435
LASI	32,656.59	13,750	2.61	92
SHARE	22,318.6	18,000	1.44	9,836
TILDA	30,899.19	25,000	1.04	2,421

Notes: The values are measured in the following units: HRS – nominal dollars before taxes and other deductions; ELSA – nominal pounds after taxes and other deductions; CRELES – 1,000 Costa Rican colons; LASI – Indian rupees; TILDA – euros; SHARE – euros, except for Denmark, Sweden, Switzerland, Poland, Czech Republic, Hungary, and Estonia. The first wave of SHARE includes before-tax income, whereas all subsequent waves consider after-tax income.

Table 3. Descriptive Statistics for Spousal Employment Earnings for the Different Data Sets.

SIEARN	Mean	Median	Skewness	Obs.
HRS	38,561.26	27,000	31.91	63,085
ELSA	15,047.19	12,309.47	13.65	16,229
CRELES	3,696.734	2,400	3.86	1,390
LASI	36,030.64	15,000	2.40	70
SHARE	22,827.07	18,511.64	1.43	6,488
TILDA	32,301.43	26,000	0.977	1,389

Notes: The values are measured in the following units: HRS – nominal dollars before taxes and other deductions; ELSA – nominal pounds after taxes and other deductions; CRELES – 1,000 Costa Rican colons; LASI – Indian rupees; TILDA – euros; SHARE – euros, except for Denmark, Sweden, Switzerland, Poland, Czech Republic, Hungary, and Estonia. The first wave of SHARE includes before-tax income, whereas all u subsequent waves consider after-tax income.

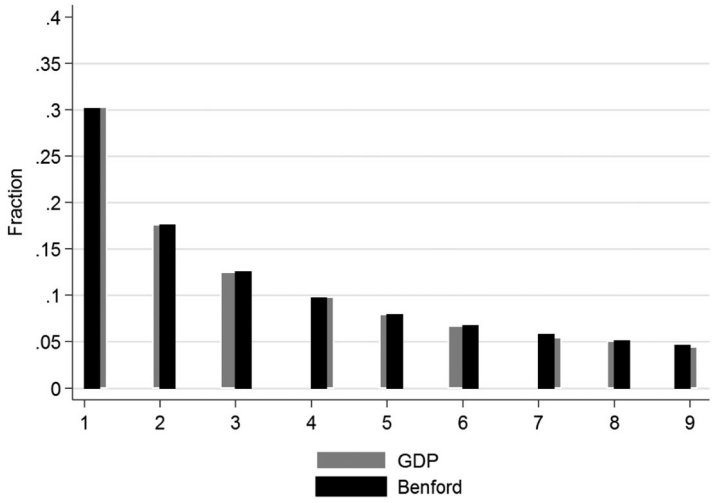

Figure 2. Relative Frequencies for GDP and Benford.

3. Results

First, to verify the assumption that income measures should generally be Benford distributed, we also include World Bank (2017) GDP data from 1960 onward measured in current U.S. dollars for 264 countries. The total number of observations is 11,315, with a mean and median of US$ 966 billion and US$ 13.8 billion, respectively. This higher mean than median, combined with the positively skewed (8.44) GDP distribution, strongly suggests that the GDP data should obey Benford's law.

By comparing the relative frequencies for the first digits in the GDP data set with those proposed by Benford's law (Figure 2), we reveal a relatively good fit, with only minor deviations. This finding is confirmed by the fact that neither the chi-squared nor Kolmogorov–Smirnov tests exceed their respective critical values (see Table 4), leading to acceptance of the null hypothesis. Likewise, the maximum deviation in the distance test is 0.0033, which corresponds to a 7% maximum (occurring at the ninth digit), a negligible discrepancy. Both these results provide strong evidence for the assumption that income data obey Benford's law, with any deviation merely an indicator of a reliability issue in the underlying data.

We then compare the frequency graph for HITOT in the harmonized data sets with that for the Benford distribution (Figure 3), revealing clearly that the first digit distribution for the income variable in the

Table 4. Test Statistic Values for GDP.

GDP	
χ^2	2.9737
D_n	0.0046
m	0.0033
n	11,315

Notes: This table lists the test statistics values for the chi-squared (χ^2), the Kolmogorov–Smirnov (D_n), and distance (m) tests. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

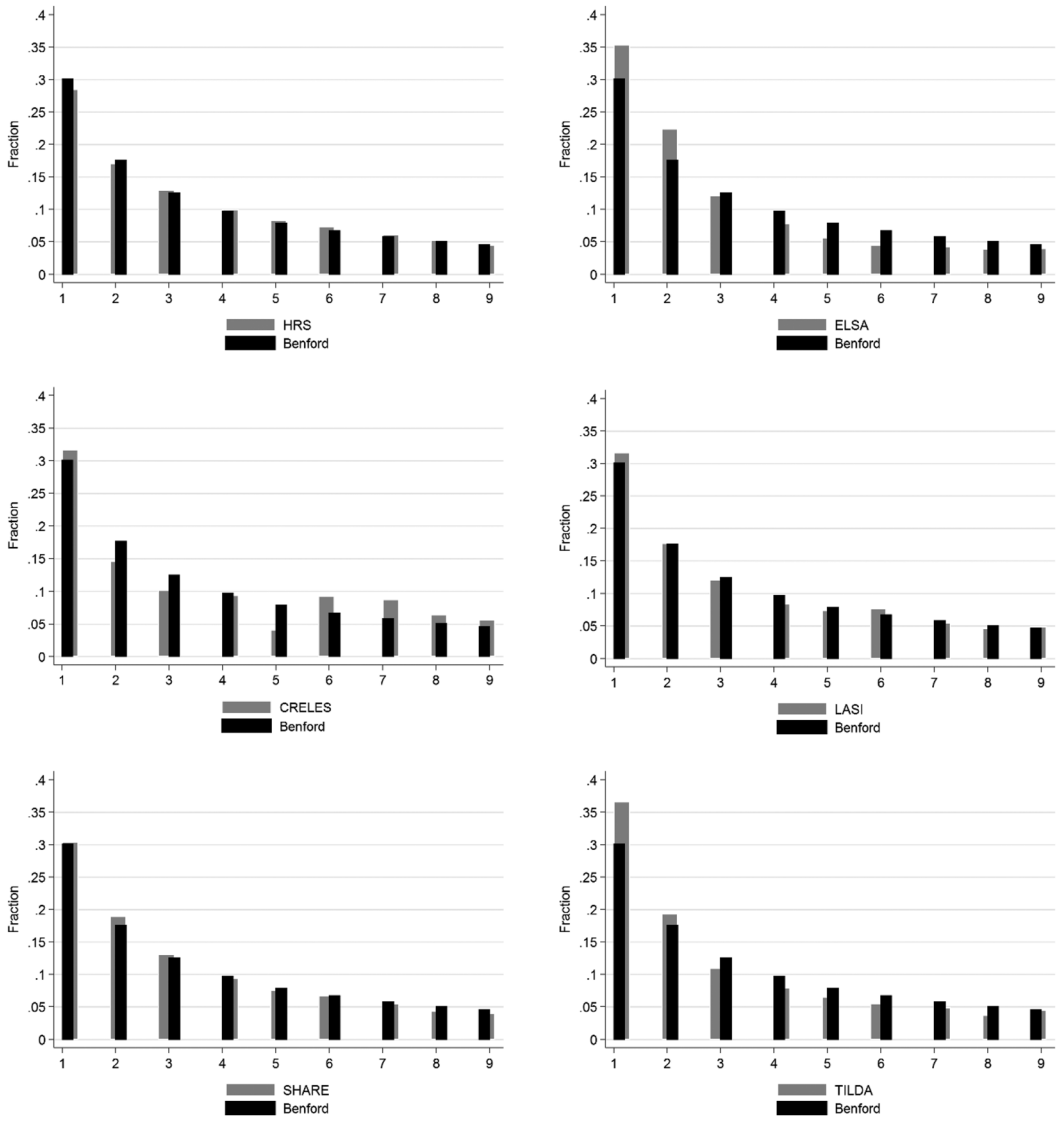

Figure 3. Relative Frequencies for HITOT and Benford for the Different Data Sets.

SHARE, HRS, and LASI data sets is close to that of the Benford distribution. The income data from the harmonized ELSA and CRELES, in contrast, although still showing a general pattern, perform comparatively poorly in terms of Benford similarity, whereas the harmonized TILDA, although generally patterned close to Benford, shows a major discrepancy in the first number.

The results revealed by this graphic analysis are largely confirmed by the statistical analysis (see Table 5). Although the HITOT variable seems to fully obey Benford's law only in the LASI data set (in which the null hypothesis cannot be rejected for either the chi-squared or Kolmogorov–Smirnov test),

Table 5. Test Statistics Values for the Different Data Sets and Variables.

	Variable											
	HITOT				*RIEARN*				*SIEARN*			
	χ^2	D_n	m	n	χ^2	D_n	m	n	χ^2	D_n	m	n
HRS	510.2509***	0.0211***	0.0159	224,285	2,314.7734***	0.0479***	0.0479	84,375	1,735.7136***	0.0511***	0.0511	63,085
ELSA	2,904.6676***	0.1001***	0.0523	61,742	1,375.7038***	0.1135***	0.099	19,796	1,212.4607***	0.1228***	0.1034	16,229
CRELES	630.3626***	0.0783***	0.0380	10,703	174.0999***	0.0468***	0.0582	2,435	148.0976***	0.0621***	0.0647	1,390
LASI	8.2094	0.0162	0.0154	1,504	8.0090	0.0902	0.0902	92	5.2130	0.0361	0.0636	70
SHARE	90.1389***	0.0223***	0.0136	25,028	518.7483***	0.0828***	0.0557	9,836	322.6380***	0.0838***	0.0503	6,488
TILDA	383.6861***	0.0821***	0.0649	12,579	85.8518***	0.0527***	0.0527	2,421	55.6337***	0.0548***	0.0548	1,389

Notes: This table lists the test statistics values for the chi-squared (χ^2), Kolmogorov-Smirnov (D_n), and distance (m) tests. The LASI results for RIEARN and SIEARN must be treated with caution because of the very small number of observations. The null hypothesis is rejected if the test statistics exceed their respective critical value – indicated by asterisks. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

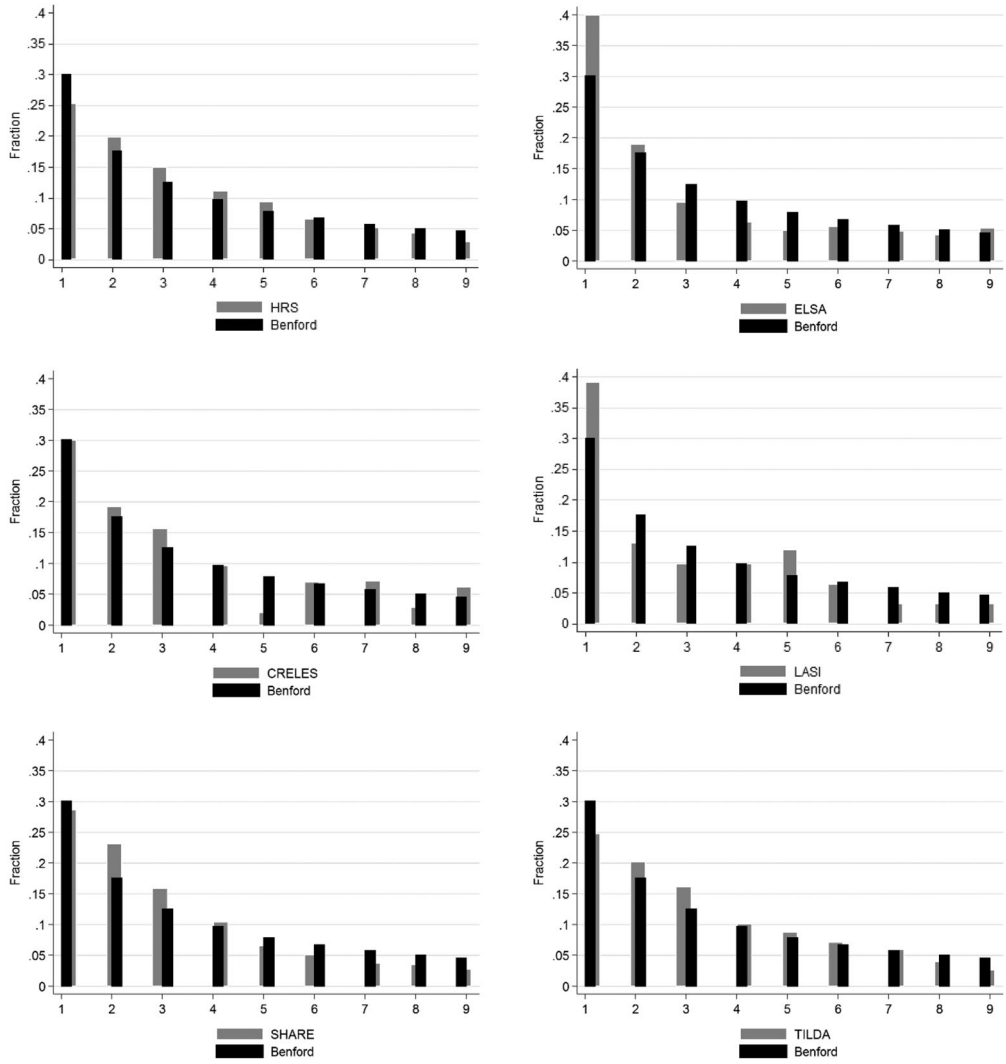

Figure 4. Relative Frequencies for RIEARN and Benford for the Different Data Sets.

the deviations indicated by the chi-square values are lower for the HRS and SHARE data than for the ELSA and CRELES data. The chi-square for the TILDA analysis seems comparatively low. When the values for each maximum deviation are compared, however, the accuracy pattern implied by the graphics holds true: the values for the ELSA, TILDA, and CRELES are higher than those for the other data sets. Hence, taken together, the graphic and empirical analyses point to major differences in the reliability of the HITOT variable across the different data sets, with particularly poor performance in the ELSA, TILDA, and CRELES data sets.

As regards RIEARN and SIEAR (Figures 4 and 5, respectively), the graphic analysis suggests that the first digits of these earnings variables do not fit Benford's law as well as does household income.

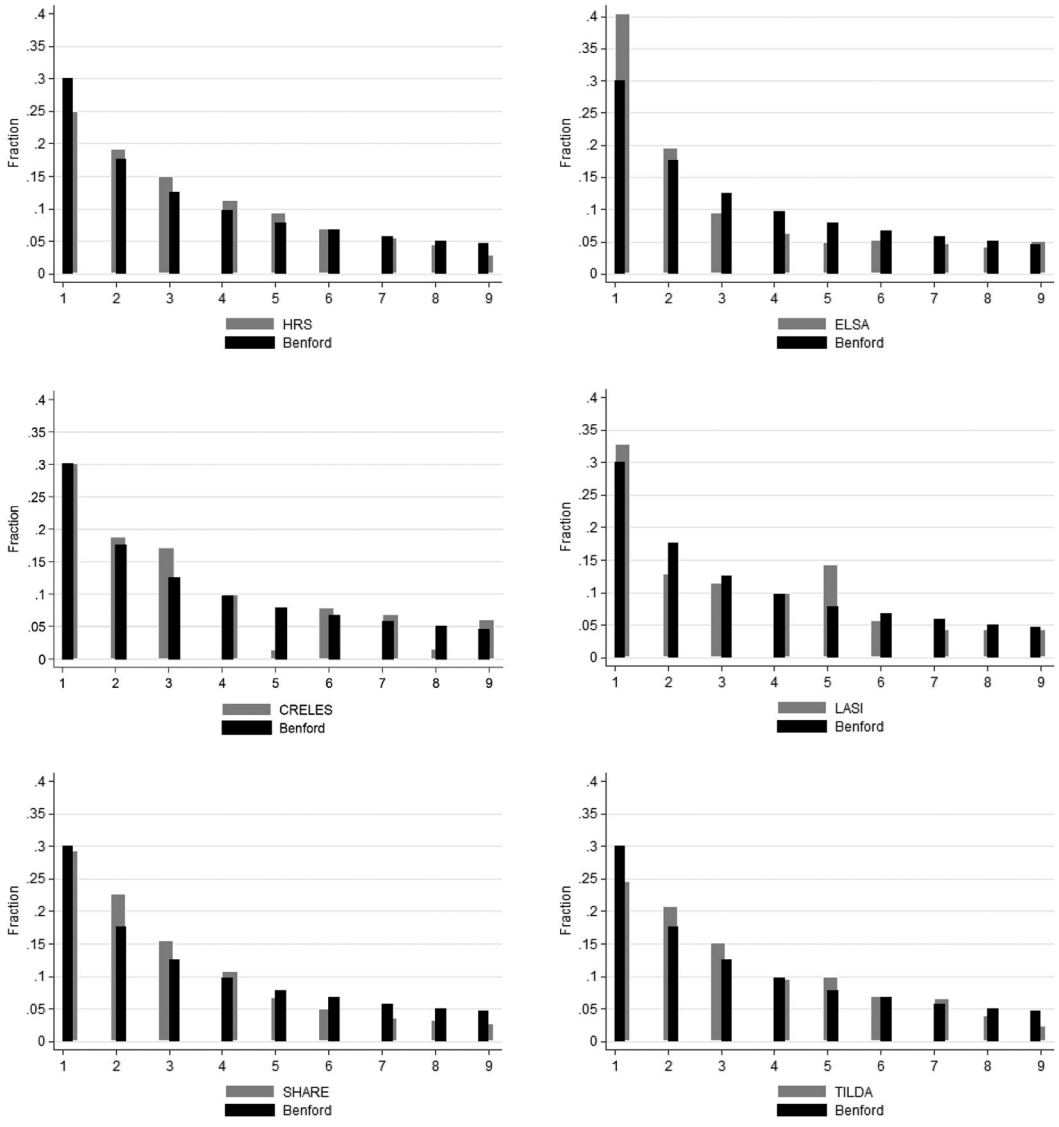

Figure 5. Relative Frequencies for SIEARN and Benford for the Different Data Sets.

Although the generally decreasing probability pattern is still present, all data sets show more or fewer major deviations for both variables. This observation is supported by the statistical analysis in which both the chi-square and Kolmogorov–Smirnov values (see Table 5) show statistically significant differences from the Benford for the first digit distributions of RIEAN and SIEARN in all data sets except the LASI. In this latter, however, acceptance of the null is mainly attributable to the low number of observations and should thus be treated with caution, as should the comparatively low chi-square values for the CRELES and TILDA data. This caveat is supported by the overall occurrence in the distance measure (for all data

sets except the TILDA) of much larger values than observed in the HITOT analysis. Given the results of both the statistical and graphic analyses, a general reliability issue with the RIEARN and SIEARN variables across all data sets analyzed is highly probable.

4. Simulation

Although the previous results raise doubts about the reliability of income data in longitudinal panel studies, these doubts are based on the assumption that income data should generally obey Benford's law. If the opposite were true, our analytic outcomes would inevitably lead to rejection of the null hypothesis because the first digit frequency patterns in the data sets analyzed would necessarily differ from those proposed by Benford. In that case, it would be impossible to use Benford's law for any valid assessment of data reliability. We thus avoid false inference by employing a simple Monte Carlo (MC) simulation in which we exploit information about the first and second moments of the HITOT variable in all our data sets to construct hypothetical samples, which are then tested for Benford adherence.[12]

The MC simulation is based on two assumptions:[13]

Assumption 1. *HITOT* $= X = (X_1, X_2, \ldots X_n)$ *is a vector of log normally distributed random variables* $log_{10}(X) \sim N(\mu_n, \sigma_n^2)$, *whose mean and variance are given by* μ_n *and* σ_n^2, *respectively.*

Assumption 2. $Y = (Y_1, Y_2, \ldots Y_n)$ *is a vector of log normally distributed variables* $log_{10}(Y) \sim N(\mu_n, \sigma_n^2)$, *whose mean and variance are given by* μ_n *and* σ_n^2, *respectively.*

The simulation itself can be characterized by the following repetitive process:

MC simulation pseudo-code:

(1) Create a vector T, with i elements
(2) Calculate $E(X) = \mu_n$ and $Var(X) = \sigma_n^2$
(3) Create new random vector Y^i containing n random variables
(4) Create vector Z^i, with $Z_n{}^i = 10^{y_n}$
(5) Conduct a Benford analysis and replace the ith element of T with the computed value for the χ^2 test statistic
(6) Repeat steps 3 to 5 i times

As Formann (2010) shows, log normally distributed random variables obey Benford's law if the probability density function fulfills certain criteria that are dependent on the values of the first two moments. Hence, the distribution of the simulated chi-square values for each data set should contain information about whether the data set should generally obey Benford's law.

In Figure 6, we show the distribution of T calculated independently for each data set when $i = 100.000$, with a vertical line indicating the critical value for the 99% confidence interval of the chi-squared distribution. Except for the ELSA data set, most simulated chi-square values are below the critical value,[14] which reinforces the assumption that the Benford deviations reported in the Results section are, in fact, the consequence of a reliability issue in the income data. It must nevertheless be noted that this finding is unconformable for ELSA, which impedes the interpretation of reliability for that particular data set.

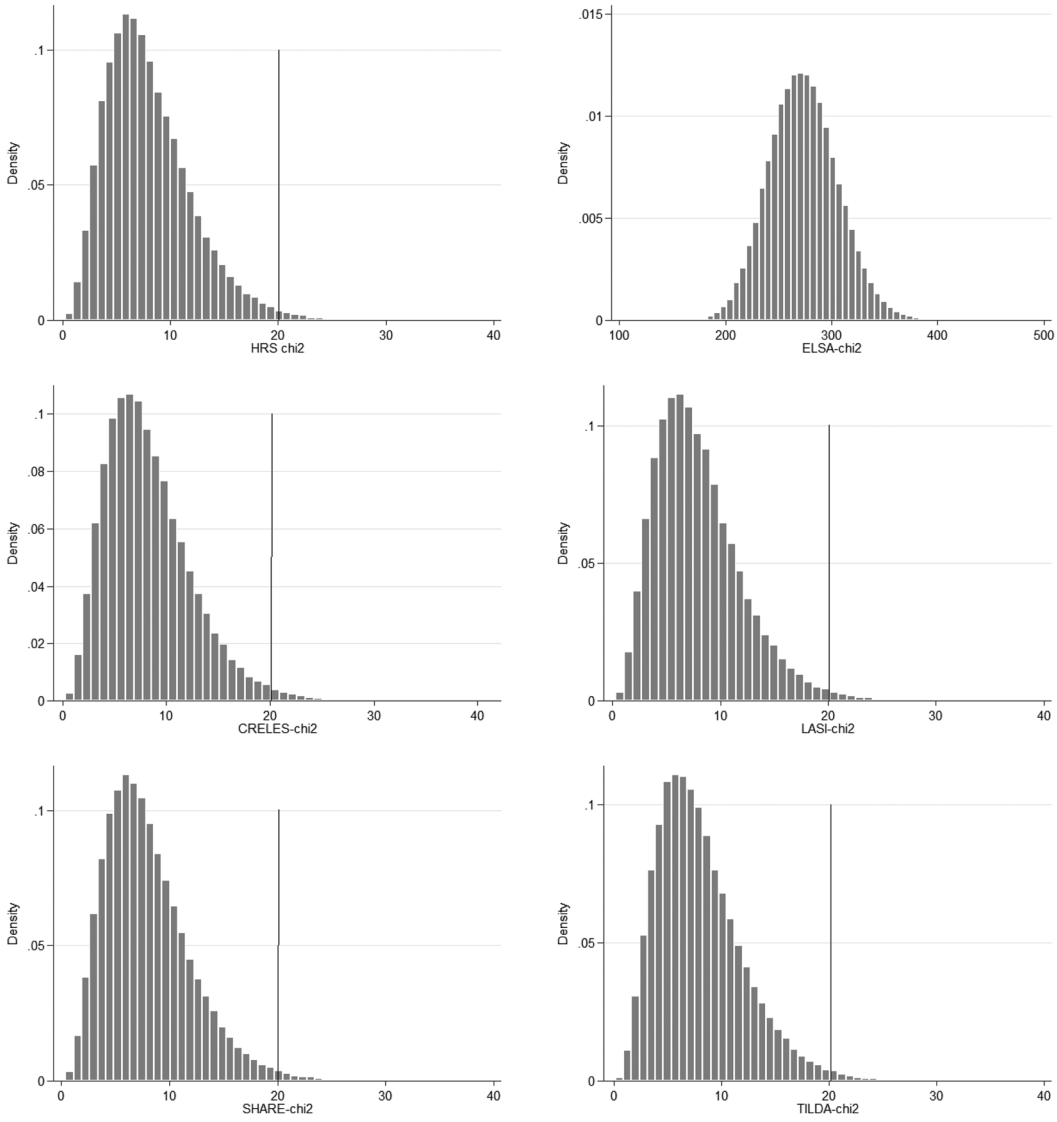

Figure 6. MC Simulation Results for Each Data Set.
Notes: The figure graphs the simulated χ^2values for each data set, with the vertical lines indicating the critical value of the chi-squared distribution for the 99% confidence interval.

5. Conclusions

According to this analysis, Benford's law seems to hold for income information from both aggregate (GDP) and survey data. Nevertheless, although there is no significant difference between the Benford first digit distribution and the first digit pattern in the World Bank GDP data, HRS data adherence to

Benford's law is less strict when measured statistically. Given income data's general tendency to conform with Benford's law, this finding strongly suggests that the income data from our surveys are subject to observational errors and should be treated with caution in terms of policy and econometric analysis. Such caution need not apply, however, to all data sets and variables analyzed. Whereas respondent (or spousal) personal earnings show poor reliability in all data sets, the household level income measure performs comparably well in the HRS (America), SHARE (Europe), and LASI (India) data.[15]

To some extent, this finding contradicts Judge and Schechter's (2009) conclusion that survey data from developing countries is generally less reliable. Rather, our analysis of harmonized data sets suggests that it is less a matter of origin and preparation than the framing of the question that determines the degree of reliability. In particular, respondents tend to make less reliable statements about their individual income than about household income.[16] The overall robustness of these findings in all but the ELSA data set is confirmed by a simulation implemented to determine whether Benford deviations do indeed indicate a reliability problem or are merely the result of a non-Benford distributed variable.

Our study results have two important implications for both econometric and policy analyses of survey data: First, our evidence of crucial reliability problems in data measured on the individual level strongly suggests the use of household level data whenever possible, second, contrary to the accepted wisdom that survey data from developed countries perform better than those from developing countries, our results suggest that any data analyzed should be evaluated for quality using a Benford analysis confirmed via simulation. As illustrated by our simulation results, this technique provides valuable insights on the variable of interest's general behavior, thereby improving the quality assessment of the underlying data. Nonetheless, because we only analyze income's Benford adherence for the first digit in a particular family of surveys, our results—whether for different variables, different survey data sets, or for second (or later) digits—are not generalizable to the broader body of survey research. Future research might thus make use of the Benford analysis to evaluate additional variables, later digits, and data from other surveys to enhance knowledge about the interrelation of data framing, origin, and reliability.

Compliance with Ethical Standards
Funding: None.
The author declares that he has no conflict of interest.
Ethical approval: This paper does not contain any studies with human participants performed by the author.

Data Availability Statement

This analysis is based primarily on income data from the harmonized datasets and codebooks developed by Gateway to Global Aging, with funding help from the National Institute on Ageing (R01 AG030153, RC2 AG036619, R03 AG043052; see www.g2aging.org). Additionally, GDP data are taken from the World Bank
(http://data.worldbank.org/indicator/NY.GDP.MKTP.CD).

Notes

1. Survey data tend to be subject to two different kinds of (nonsampling) measurement errors: random errors, such as those caused by interviewer inattentiveness, or nonrandom (systematic) errors, such as the inaccurate responses generated when the survey is badly constructed or includes problematic survey items (for a detailed discussion, see Groves, 2004; Saris and Gallhofer, 2014; Bowling, 2005)
2. Saris and Revilla (2016) provide a useful overview of existing correction techniques for measurement errors in survey data. Moreover, a wide body of literature focuses on measurement error in time-series

data. Such techniques usually exploit the data's distributional properties to overcome the uncertainty from measurement problems (e.g., the Kalman (1960) filter, a powerful tool for improving predictions in an evolving system).

3. For instance, a Benford analysis facilitated the uncovering of the 2001 Enron accounting fraud (Nigrini, 2012, p. 207). See also Ausloos, Cerqueti and Mir (2017) for a more recent study about tax fraud and Benford´s law.
4. A more generalized formulation of Benford's law describes the probability of the occurrence of a particular number j as the nth digit in the following form: $p(j) = \sum_{k=B^{n-1}}^{B^n-1} log_B(1 + \frac{1}{k \cdot B+j})$. Readers interested in the law's application beyond the first digits addressed here will find additional information in Hill (1995b) and Durtschi *et al.* (2004).
5. Here, $\Pr_n(j)$equals the probability of observing a number beginning with digit j in a given data set with n observations (i.e., the relative frequency of numbers beginning with j in a given data set).
6. For more information, see https://g2aging.org/?.
7. The Department of Health, the Department of Work and Pensions, and the Department for Transport.
8. SHARE data cover the following European countries: Austria, Belgium, Czech Republic, Denmark, Estonia, France, Germany, Greece, Hungary, Ireland, Italy, Luxembourg, the Netherlands, Poland, Portugal, Slovenia, Spain, Sweden, and Switzerland.
9. Minor differences in income variable composition among the different surveys include the inclusion (HRS) or exclusion (TILDA) of second job earnings in constructing the RIEARN and SIEARN variables (see the respective data set codebooks for more information).
10. For instance, although almost all harmonized data sets include information about drinking behavior, some questionnaires ask respondents for their daily alcohol intake (e.g., ELSA or LASI), whereas others ask for the total number of drinks if the respondent is currently drinking (e.g., TILDA).
11. For example, the German Federal Ministry of Labor and Social Affairs regularly publishes a Poverty and Wealth Report (*Armuts-und Reichtumsbericht*), which deals with income dynamics among Germans. The report's major analyses and conclusions are based on income data from the German Socio-Economic-Panel (GSOEP), a large-scale, national, longitudinal survey (see http://www.armuts-und-reichtumsbericht.de/DE/Startseite/start.html).
12. We apply the MC simulation only to the HITOT variable because the calculations are so computationally intensive.
13. We base Assumption 1 on the demonstrable tendency of income data to be asymptotically log normally distributed (Clementi and Gallegati, 2005).
14. The 99th percentile values are 20.0644 (HRS), 354.0885 (ELSA), 20.0874 (CRELES), 20.0644 (LASI), 20.1608 (SHARE), and 20.2041 (TILDA).
15. However, since there exists no clear cutoff for each statistical measure that indicates a change from a good to a rather poor reliability, it is rather the relative discrepancy in the test statistics for different data sets that matters in the assessment of the quality. For instance, while the SHARE data show a chi-square value of around 90 for the analysis of HITOT, the tests statistic is 7 times higher for the CRELES data. Given this discrepancy—together with the results of the Kolmogorov–Smirnov as well as the distance measure—the SHARE shows a rather high quality compared to the CRELES data.
16. Because all the data sets employ detailed income measures, the differences in these measures' reliability do not stem from rounding errors.

References

Ausloos, M., Herteliu, C. and Ileanu, B. (2015) Breakdown of Benford's law for birth data. *Physica A: Statistical Mechanics and Its Applications* 419: 736–745. http://doi.org/10.1016/j.physa.2014.10.041

Ausloos, M., Cerqueti, R. and Mir, T.A. (2017) Data science for assessing possible tax income manipulation: The case of Italy. *Chaos, Solitons & Fractals* 104: 238–256. https://doi.org/10.1016/j.chaos.2017.08.012

Berger, A., Hill, T.P., Kaynar, B. and Ridder, A. (2011) Finite-state Markov Chains Obey Benford's Law. *SIAM Journal on Matrix Analysis and Applications* 32(3): 665–684.

Bowling, A. (2005) Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health* 27(3): 281–291. http://doi.org/10.1093/pubmed/fdi031

Clementi, F. and Gallegati, M. (2005) Pareto's Law of Income Distribution: Evidence for Germany, the United Kingdom, and the United States. In A. Chatterjee, S. Yarlagadda and B.K. Chakrabarti (eds.), *Econophysics of Wealth Distributions: Econophys-Kolkata I* (pp. 3–14). Milano: Springer Milan. http://doi.org/10.1007/88-470-0389-X_1

Deleanu, I.S. (2017) Do countries consistently engage in misinforming the international communityabout their efforts to combat money laundering? Evidence using Benford'slaw. *Plos One* 12(1). https://doi.org/10.1371/journal.pone.0169632

Duncan, G.J. and Hill, D.H. (1985) An investigation of the extent and consequences of measurement error in Labor-Economic Survey Data. *Journal of Labor Economics* 3(14): 508–532.

Durtschi, C., Hillison, W. and Pacini, C. (2004) The effective use of Benford's law to assist in detecting fraud in accounting data. *Journal of Forensic Accounting* 99(99): 17–34.

Formann, A.K. (2010) The Newcomb–Benford law in its relation to some common distributions. *PLoS One* 5(5): e10541. http://doi.org/10.1371/journal.pone.0010541

Fu, D., Shi, Y.Q. and Su, W. (2007) A generalized Benford's law for JPEG coefficients and its applications in image forensics. *Security, Steganography, and Watermarking of Multimedia Contents IX* (Vol. 6505). International Society for Optics and Photonics.

[Dataset] Gateway to Global Aging with funding help from the National Institute on Ageing (2017) *Harmonized Datasets* [Data file]. www.g2aging.org

Gauvrit, N.G., Houillon, J.C. and Delahaye, J.P. (2017) Generalized Benford's Law as a lie detector. *Advances in cognitive psychology* 13(2): 121. https://doi.org/10.5709/acp-0212-x

Groves, R.M. (2004) *Survey errors and survey costs*. Hoboken, New Jersey: John Wiley & Sons.

Hill, T.P. (1995) A statistical derivation of the significant-digit law. *Statistical Science* 10(4): 354–363. http://doi.org/10.2307/2246134

Hill, T.P. (1995a) Base-invariance implies benford's law. *Proceedings of the American Mathematical Society* 123(3): 887–895.

Hill, T.P. (1995b) The signifilcant digit phenomenon. *The American Mathematical Monthly* 102(4): 322–327.

Judge, G. and Schechter, L. (2009) Detecting problems in survey data using Benford's law. *Journal of Human Resources* 44(1): 1–24. http://doi.org/10.1353/jhr.2009.0010

Kalman, R.E. (1960) A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* 82(1): 35–45.

Le Gallo, J. (2004) Space-time analysis of GDP disparities among European regions: A Markov chains approach. *International Regional Science Review* 27(2): 138–163.

Leemis, L.M.L., Schmeiser, B.W.S. and Evans, D.L.E. (2000) Survival distributions satisfying Benford's law. *The American Statistician* 54(3): 236–241.

Mellow, W. and Sider, H. (1983) Accuracy of response in labor market surveys: Evidence and implications. *Journal of Labor Economics* 1(4): 331–344.

Mir, T.A. (2014) The Benford law behavior of the religious activity data. *Physica A: Statistical Mechanics and Its Applications* 408: 1–9. http://doi.org/10.1016/j.physa.2014.03.074

Nigrini, M. (2012) *Benford's law: Applications for forensic accounting, auditing, and fraud detection* (Vol. 586). Hoboken, NJ: John Wiley & Sons.

Nigrini, M.J. and Miller, S.J. (2007) Benford's Law applied to hydrology data—Results and relevance to other geophysical data. *Mathematical Geology* 39(5): 469–490. http://doi.org/10.1007/s11004-007-9109-5

Pietronero, L., Tosatti, E., Tosatti, V. and Vespignani, A. (2001) Explaining the uneven distribution of numbers in nature: The laws of Benford and Zipf. *Physica A: Statistical Mechanics and Its Applications* 293(1–2): 297–304. http://doi.org/10.1016/S0378-4371(00)00633-6

Pinkham, R.S. (1961) On the Distribution of First Significant Digits. *The Annals of Mathematical Statistics*. http://doi.org/10.2307/3615826

Sachs, L. (2004) *Angewandte Statistik*. Berlin, Heidelberg: Springer.
Sandron, F. (2002) Les populations suivent-elles la loi des nombres anomaux? *Population* 57(4): 761–767. http://doi.org/10.2307/1534803
Saris, W.E. and Gallhofer, I.N. (2014) *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken, NJ: John Wiley & Sons. http://doi.org/http://doi.org/10.1002/9780470165195
Saris, W.E. and Revilla, M. (2016) Correction for measurement errors in survey research: Necessary and possible. *Social Indicators Research* 127(3): 1005–1020. http://doi.org/10.1007/s11205-015-1002-x
Shi, J., Ausloos, M. and Zhu, T. (2018) Benford's law first significant digit and distribution distances for testing the reliability of financial reports in developing countries. *Physica A: Statistical Mechanics and its Applications* 492: 878–888. https://doi.org/10.1016/j.physa.2017.11.017
Smirnov, N. (1948) Table for Estimating the Goodness of Fit of Empirical Distributions. *Annals of Mathematical Statistics* 19(2): 279–281. http://doi.org/10.1214/aoms/1177730256
Swanson, D., Cho, M. and Eltinge, J. (2003) Detecting possibly fraudulent or error-prone survey data using Benford's Law. *Proceedings of the Section on Survey Research Methods, American Statistical Association* (pp. 937–941).
Tödter, K.H. (2009) Benford's law as an indicator of fraud in economics. *German Economic Review* 10(3): 339–351. http://doi.org/10.1111/j.1468-0475.2009.00475.x
Villas-Boas, S.B., Fu, Q. and Judge, G. (2017) Benford's law and the FSD distribution of economic behavioral micro data. *Physica A: Statistical Mechanics and its Applications* 486: 711–719.
Wallace, W.A. (2002) Assessing the quality of data used for benchmarking and decision-making. *The Journal of Government Financial Management* 51(3): 16–22.
Whyman, G., Ohtori, N., Shulzinger, E. and Bormashenko, E. (2016) Revisiting the Benford law: When the Benford-like distribution of leading digits in sets of numerical data is expectable?. *Physica A: Statistical Mechanics and its Applications* 461: 595–601.
[Dataset] World Bank, World Bank Open Data (2017) *GDP (current US$)* [Data file]. https://data.worldbank.org/indicator/NY.GDP.MKTP.CD