Empirica (2009) 36:273–292
DOI 10.1007/s10663-008-9084-1

ORIGINAL PAPER

Does Benford's Law hold in economic research and forecasting?

Stefan Günnel · Karl-Heinz Tödter

Published online: 27 August 2008
© Springer Science+Business Media, LLC. 2008

Abstract First and higher order digits in data sets of natural and socio-economic processes often follow a distribution called Benford's law. This phenomenon has been used in business and scientific applications, especially in fraud detection for financial data. In this paper, we analyse whether Benford's law holds in economic research and forecasting. First, we examine the distribution of regression coefficients and standard errors in research papers, published in *Empirica* and *Applied Economics Letters*. Second, we analyse forecasts of GDP growth and CPI inflation in Germany, published in *Consensus Forecasts*. There are two main findings: The relative frequencies of the first and second digits in economic research are broadly consistent with Benford's law. In sharp contrast, the second digits of *Consensus Forecasts* exhibit a massive excess of zeros and fives, raising doubts on their information content.

Keywords Benford's Law · Fraud detection · Regression coefficients · Standard errors · Growth and inflation forecasts · Rounding

JEL Classification C8 · C52 · C12

The paper was written while the first author was a visiting researcher at the Research Centre of the Deutsche Bundesbank. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Deutsche Bundesbank. We thank two anonymous referees and the editor for helpful comments and suggestions.

S. Günnel · K.-H. Tödter (✉)
Research Centre Deutsche Bundesbank, Wilhelm-Epstein-Strasse 14, 60431 Frankfurt am Main, Germany
e-mail: karl-heinz.toedter@bundesbank.de

1 Introduction

Increasing differentiation and growing social and economic relevance of research raises the temptation to make up research results (Reulecke 2006). This process is fuelled by increasing publication pressure in academics. The traditional control mechanisms in the publication process, such as anonymous refereeing, are easily overstrained in dealing with empirical research papers using large data sets and complex econometric tools.

However, independent reviews of the outcome of empirical research are a cornerstone of science (Hamermesh 2007). In contrast to natural sciences, there is no distinct tradition of replication in social sciences. In economics, most academic journals do not request from their authors the filing of data and programs.[1] But "research that cannot be replicated is not science, and cannot be trusted either as part of the profession's accumulated body of knowledge or as a basis for policy." (McCullough and Vinod 2003, p. 888) Thus, if the outcome of empirical research in economics is replicable by independent experts only in rare cases, indirect tools for detecting fraud and checking reliability are called for.

Already in 1972, U.S. economist Hal Varian proposed to use Benford's law as a diagnostic tool for screening model output, in particular forecasts, for irregularities that deserve closer inspection. In many data sets, from newspaper articles to the length of rivers, Benford's law has been found to hold surprisingly well. More recently, Benford's law has been applied quite successfully to detect fraud and manipulation in business and administration data like balance sheets and tax declarations. Moreover, experimental research has shown that people are not particularly good at replicating known pattern of data. For instance, they tend to over-report modes and to avoid long runs (Camerer 2003, p. 134). Benford's law, though widely applicable, is not yet widely known. Since it is unlikely that manipulated numbers would preserve it, Benford's law is a potentially useful diagnostic. Diekmann (2007) investigated sociological empirical research, testing regression coefficients and other statistics for deviations from Benford's law. To our knowledge, in the field of economics tests of Benford's law have not yet been applied to published empirical research and forecast data.

Regression results are a major outcome of empirical economic research and published economic forecasts are an important source of information in the decision making process of economic agents, including governments and central banks. This paper investigates empirically whether Benford's law can serve as a tool for detecting irregularities in economic research and forecasting that may deserve closer scrutiny. After a brief introduction to Benford's law, Section 2 reviews some aspects of fraud detection with Benford's law. Section 3 applies Benford's law to test econometric research published in *Empirica* and *Applied Economics Letters*. Section 4 examines GDP growth and CPI inflation forecasts for Germany published in *Consensus Forecasts*. Section 5 concludes.

[1] Even if that is the case, attempts to replicate the studies mostly fail. McCullough et al. (2006) analysed more than 150 articles from the *Journal of Money, Credit, and Banking*, but were able to reproduce the results in less than 10 percent of the cases.

1.1 What is Benford's law?

Intuitively, one may think that the first digits of numbers are uniformly distributed, i.e. numbers are equally likely to start with 1, 2 or 9. The American astronomer Simon Newcomb (1881) observed that the first pages of logarithmic tables (containing numbers beginning with 1, 2, 3) were more worn out than the last pages (numbers starting with 7, 8, 9). He concluded that lower digits seem to appear more often than higher digits. Zero as a first digit is ignored. Newcomb calculated relative frequencies of the first (d_1) and second (d_2) significant digits according to the formulas:

$$p(d_1) = \log_{10}\left(1 + \frac{1}{d_1}\right) \qquad d_1 = 1, 2, \ldots, 9 \tag{1}$$

$$p(d_2) = \sum_{k=1}^{9} \log_{10}\left(1 + \frac{1}{10k + d_2}\right) \qquad d_2 = 0, 1, 2, \ldots, 9 \tag{2}$$

However, Newcomb's findings were forgotten until the American General Electric physicist Frank Benford (1938) rediscovered the first digit phenomenon. Benford analysed 20 data sets including population statistics, figures published in newspapers, American League baseball statistics, atomic weights of chemical elements etc. with more than 20,000 first digits in total. Hill (1995) derived the joint distribution of the first and higher-order significant digits:

$$p(D_1 = d_1, \ldots, D_k = d_k) = \log_{10}\left(1 + \left(\sum_{i=1}^{k} d_i 10^{k-j}\right)^{-1}\right) \tag{3}$$

$$\forall k \in Z, \quad d_1 \in \{1, 2, \ldots, 9\} \quad \text{and} \quad d_j \in \{0, 1, 2, \ldots, 9\}, j = 2, \ldots k$$

Applying this formula to a combination of digits, e.g. 25, yields $p(D_1 = 2,\ D_2 = 5) = \log_{10}\left(1 + (25)^{-1}\right) \cong 0.017$. Table 1 displays the joint probabilities for combinations of the first two digits. The marginal probabilities of the first and the second digits are shown in the final column and row, respectively.

Table 1 Benford distribution

d1/d2	0	1	2	3	4	5	6	7	8	9	p(d1)
1	0.041	0.038	0.035	0.032	0.030	0.028	0.026	0.025	0.023	0.022	0.301
2	0.021	0.020	0.019	0.018	0.018	0.017	0.016	0.016	0.015	0.015	0.176
3	0.014	0.014	0.013	0.013	0.013	0.012	0.012	0.012	0.011	0.011	0.125
4	0.011	0.010	0.010	0.010	0.010	0.010	0.009	0.009	0.009	0.009	0.097
5	0.009	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.007	0.007	0.079
6	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.006	0.006	0.006	0.067
7	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.005	0.058
8	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.051
9	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.004	0.004	0.004	0.046
p(d2)	0.120	0.114	0.109	0.104	0.100	0.097	0.093	0.090	0.088	0.085	1.000

Pinkham (1961) and Hill (1995) proved that Benford's law is base invariant (i.e. the distribution remains unchanged irrespective whether the numbers are expressed in base 2, 4, 8 etc.) and scale invariant (e.g. if Benford's law holds for distances expressed in kilometres, it also holds if the data are transformed into miles). However, as will be discussed later, Benford's law is not invariant to rounding. Mathematical explanations for the appearance of Benford's law can be found in Hill (1995, 1998) who proved a "random samples from random distributions theorem." It states, under fairly general conditions, that if distributions are selected randomly and random samples are taken from each distribution, then the frequency of digits will converge to Benford's law.

2 Detecting fraud with Benford's law

2.1 Brief review of applications in business and economics

In the last two decades, in particular, Benford's law was increasingly applied to business and scientific data as a method to identify fraud or manipulation. Recently, Diekmann (2007) investigated the first and second digits of published statistical results in the field of sociology. He analysed regression results in two samples (approximately 2,600 observations) drawn from four volumes of the *American Journal of Sociology.* In addition, he investigated results from experimental studies with test persons. While he found that (the first and second digits of) published regression coefficients approximately obey Benford's law, his experimental data suggest that the second digits (not the first ones) of faked regression coefficients were less in accordance with it.

Carslaw (1988) investigated the second digits of profits of New Zealand firms and found that managers tend to round up the firm's profits due to psychological reasons. A profit of € 3.00 million appears to be much higher than a profit of € 2.99 million. Thus, there is an excess of zeros but a lack of nines in the second digits compared to the Benford distribution. Similar results were found by Thomas (1989), who conducted a study for U.S. firms, distinguishing between profits and losses. While the results for U.S. firms' profits are in line with Carslaw, he finds the reverse phenomenon for losses, i.e. managers tend to optically "shrink" losses by rounding appropriately (less zeros, more nines). Additional studies on this issue have been conducted by Niskanen and Keloharju (2000) for Finnish companies and Van Caneghem (2002) for U.K. companies.

Nigrini's (1996a, b, 1999) publications were quite influential for introducing Benford's law in finance and accountancy. He analysed tax declarations of American taxpayers and figured out that people tend to understate their true taxable income. Due to U.S. law, where taxes are set after tax tables, even minor understatements can result in significant tax reductions. These findings inspired tax authorities in e.g. the U.S., Switzerland, the Netherlands and Germany, to check tax declarations for inconsistencies by applying Benford's law. Recently, Quick and Wolz (2003) examined balance sheet and income statement data of German companies for the years 1994–1998. Their results show that the first and second digits in most of the cases (on a year by year analysis as well as for the whole period) closely follow the Benford distribution.

Benford’s law has also been applied to check predictions of mathematical models for plausibility provided that the real data follow Benford’s law. Ley (1996) has shown that a series of one-day returns (using data for more than half of the 20th century) on the Dow Jones Industrial Average Index and the Standard & Poor’s Index is in line with the Benford’s distribution. A similar result is obtained by Tödter (2007) for the first digits of closing prices of German stocks. Moreover, he shows that the predictions for share prices by the Black and Scholes model are consistent with the Benford distribution for the first digit. In addition, Benford’s law can be applied to test for psychological barriers in stock markets (see De Ceuster et al. 1998 among others) and ebay auctions (Giles 2007). Schatte (1988) showed that the expected storage space for computers with binary-base is at its minimum for base 8. Recent results for survey data, e.g. the German Socio-Economic Panel, can be found in Schräpler and Wagner (2005) and Schäfer et al. (2005).

Provided the population of specific data is distributed according to Benford’s law it is widely accepted in empirical literature that manipulated data do no longer adhere to the specific distribution. However, in general one can not conclude a priori that a certain data set contains faked numbers if it deviates from Benford’s law. Hence, in a first step, it needs to be established that the Benford distribution applies to the population of a data set before an *appropriate* sample is checked for deviations.

2.2 Requirements to data sets for testing Benford’s law

Benford (1938, p. 552) stated that “the method of study consists of selecting any tabulation of data that is not too restricted in numerical range, or conditioned in some way too sharply.” More precisely, in the literature a number of “rules” are formulated (see Durtschi et al. 2004 and Mochty 2002 among others) on which data are expected to follow Benford’s law. The data set should either be complete or a random sample drawn from it to avoid biases. Moreover, data should be expressed in the same dimensions such as dollar or miles. Mochty (2002) advises not to use statistical estimates (means, variances) since they themselves follow certain distributions (Normal-, Chi^2- etc.). However, that does not preclude the leading digits to obey Benford’s law. Some of these statistics are checked in this study with surprising results. It is unanimously agreed in the literature that data shall not be restricted to certain minimum or maximum values (e.g. the body height of persons). Problems may also arise where data are restricted by psychological barriers (e.g. prices in supermarkets often have nine as a last digit—€ 1.99). In addition, numbers shall not be artificial or made up by humans (e.g. telephone numbers, postal codes). Last but not least, rounded numbers do on average no longer follow Benford’s law even if the original data do.

2.3 Testing Benford’s law

Several statistical tests can be applied to inspect whether the distribution of the first and higher order digits conforms to Benford’s law, such as the Chi^2 test, the Mean test or the Kuiper test. If h_d (p_d) denotes the observed relative frequencies (probabilities) of digit d in a data set with N observations, the Chi^2-statistics for first and second (and higher order) digits are defined as

$$T_1 = N \sum_{d_1=1}^{9} \frac{(h_{d_1} - p_{d_1})^2}{p_{d_1}}, \; T_2 = N \sum_{d_2=0}^{9} \frac{(h_{d_2} - p_{d_2})^2}{p_{d_2}} \tag{4}$$

Under the null hypothesis of the Benford law, the statistics are Chi^2—distributed with 8 (9) degrees of freedom. As quadratic measures, the statistics are sensitive to the pattern of deviations from Benford's law. Moreover, with a fixed significance level α and increasing sample size (N), the tests will eventually reject the null, as the probability of a type II error (β) approaches zero.[2]

Under Benford's law the mean of the first digit is 3.940 (with variance 6.057) and the mean of the second digit is 4.687 (with variance 8.254). To test whether the mean of the observed digits, calculated as $\bar{d}_1 = \sum_{d_1=1}^{9} (d_1 + 0.5)\, h_{d_1}$ and $\bar{d}_2 = \sum_{d_2=0}^{9} (d_2 + 0.5)\, h_{d_2}$, respectively, deviates from these values, the approximately standard normal statistics

$$T_{\bar{d}_1} = \sqrt{N}\, \frac{\bar{d}_1 - 3.940}{\sqrt{6.057}}, \quad T_{\bar{d}_2} = \sqrt{N}\, \frac{\bar{d}_2 - 4.687}{\sqrt{8.254}} \tag{5}$$

can be used. The Mean tests are less sensitive to deviations in single digits and less responsive to the sample size.[3]

The Kuiper (1959) test is a modification of the Kolmogorov–Smirnov test (Giles 2007). Let H_d (P_d) denote the cumulated empirical relative frequencies (cumulated probabilities), then the Kuiper-statistic is

$$T_K = \left(D_N^+ + D_N^-\right)\left[\sqrt{N} + 0.155 + 0.24/\sqrt{N}\right] \tag{6}$$

where $D_N^+ = \sup\,[H_d - P_d]$ and $D_N^- = \sup\,[P_d - H_d]$. [4]

3 Benford's law in published econometric research

Intuitively, if a researcher intends to manipulate regression results to confirm or to refute a specific economic hypothesis, he is most likely to forge the leading digits, i.e. the first and second digit, of estimated coefficients and/or standard errors.

[2] For specific digits, e.g. whether there is an excess of fives, the standard normal statistic $T_d = \sqrt{N_d}(h_d - p_d)/\sqrt{p_d(1 - p_d)}$ can be used to check whether the observed frequency significantly deviates from its theoretical value.

[3] A closely related statistic is Nigrini's (1996a,b) distortion factor.

[4] Recently, Tam Cho and Gaines (2007) proposed the Euclidean distance as a measure to characterize the deviation from the Benford distribution. This measure is independent of the sample size, however, it is lacking a statistical foundation.

Table 2 Test statistics for the first digits of regression coefficients

Empirica	2003	2004	2005	2006	2003–06
Number of observations N	931	643	1,352	1,680	4,606
N per article	78	58	135	129	100
Chi^2 test	32.60***	11.63	27.41***	19.08**	11.35
Probability	0.00	0.17	0.00	0.01	0.18
Kuiper test	1.83**	1.04	1.91**	1.65*	0.95
Mean test (absolute value)	1.69*	0.50	2.94***	3.05***	1.19

*, **, *** denote a significant test value on the 10%-, 5%- and 1%-level

The critical test values for the respective significance levels are as follows:

Chi^2 test (8 df): 13.36, 15.51, 20.09; Kuiper test: 1.62, 1.75, 2.00; Mean test: 1.64, 1.96, 2.58. They apply throughout the paper for any first digit analysis

Hence, the analysis focuses on the first and second digits only.[5] To test the Benford hypothesis, we investigate volumes 30, 31, 32, and 33 of *Empirica* (years 2003 to 2006) with more than 14,000 first and second digits of coefficients and standard errors. In order to check the robustness of the results, volume 13 of *Applied Economics Letters* (year 2006) with more than 15,000 observations is analysed, too.

We collected regression coefficients and standard deviations from a broad range of regression types, e.g. OLS, (inter-) quantile regressions, GMM, IV estimations, (censored) Tobit regressions, random and fixed effects estimations, SURE, VAR-models. Thereby, only regression results from empirical data are considered but no data obtained by simulation procedures.

Not in all cases a standard error (S.E.) was published along with the coefficient. If possible, the S.E. was calculated from the published t-value, taking into account that this might cause rounding problems. To illustrate this point, imagine that the original value of the coefficient is 1.394 with a t-value of 3.475. Calculating the S.E. gives 0.40115108. Suppose, the published data in an article are 1.39 and 3.48 for the coefficient and the t-value, respectively. The calculated S.E. equals 0.39942529. Obviously, this will cause misleading results for testing the digits frequencies. Keeping that in mind, we will comment on the importance of this phenomenon later.

For convenience, only regression results presented in tables of the respective journals and articles are included in the study, which is by far the majority of all available data. Moreover, in the subsequent analysis, it is not distinguished between positive and negative regression coefficients since there is no justification for doing so.

3.1 Results for first digits of regression coefficients in Empirica

We start by presenting the results for *Empirica*. The test statistics for the first digits of the regression coefficients are displayed in Table 2. Looking first at the test

[5] The results for third digits have been evaluated as well (overall showing a very good agreement with Benford's law) but are not reported due to space limitations. An analysis of higher-order digits (which are more likely to be uniformly distributed) is impeded by insufficient digits in most published papers.

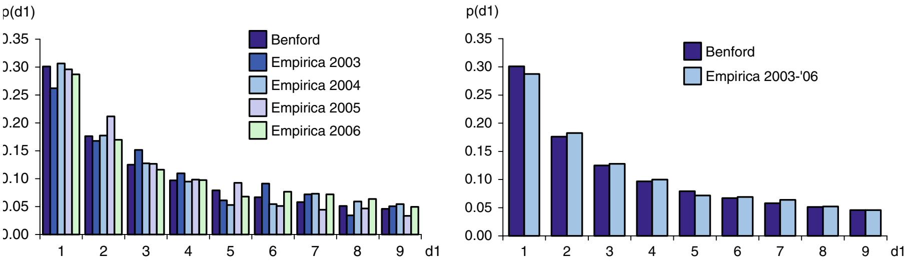


Fig. 1 Relative frequencies of the first digits of regression coefficients

statistics for the total sample, none of the tests yields a significant value even on a 10% significance level. Thus, the Benford distribution cannot be rejected.

However, results are more diverse if the individual years are examined. The respective relative frequencies are displayed in Fig. 1. The year 2004 with the smallest number of observations (N = 643) has the best statistical fit (no significance on a 10% level) to Benford's law. In contrast, the observations in *Empirica* 2006 show significant test statistics on a 10% (Kuiper test), 5% (Chi^2 test) and 1%-level (Mean test). Although graphically the fit of the 2006 data appears to be slightly better than the 2004 one does, the number of observations is much higher (N = 1,680) which boosts the test statistics towards the rejection region. Furthermore, it is worth noting that on average in 2006 there are approximately twice as much coefficients per article (~129) as in 2004 (~59). Hence, the dependency on single articles is higher. Regarding the years 2003 and 2005, the test statistics for the Chi^2 test are significant on a 1% level and on a 5% level for the Kuiper test and suggest to reject the null of a Benford distribution. Graphically, in 2003 the digits 1, 5 and 8 are under-represented whereas 3 and 6 appear too often. The dubious test statistics for 2005 can be attributed to the high relative frequency of digit two. It should be pointed out that the tendency of the Kuiper test to reject the null less frequently than the Chi^2 test has been verified in many of our samples. In contrast, the Mean test does not show such a clear tendency. For 2003 it rejects the null on a 10% level, for 2005 on a 1% level.

In summary, although three out of four sub-samples seem to reject (at least partly) Benford's law for the first significant digit of regression coefficients, this effect averages out in the total sample.

3.2 Results for second digits of regression coefficients in Empirica

The test statistics of the second digits of regression coefficients are displayed in Table 3, the graphical output in Fig. 2. It can be seen that the number of observations slightly drops compared to the first digit, because some published coefficients only have one significant digit. On average, the test statistics are more in line with Benford's law than for the first digit. Only in the total sample the null is marginally rejected at a 5% significance-level with the Chi^2 test. All other tests are insignificant on a 10% level, strongly suggesting that Benford's law applies.

Table 3 Test statistics for the second digits of regression coefficients

Empirica	2003	2004	2005	2006	2003–06
Number of observations N	831	550	1,067	1,529	3,977
N per article	69	50	107	118	86
Chi^2 test	12.15	11.47	14.36	6.68	17.99**
Probability	0.20	0.24	0.11	0.67	0.04
Kuiper test	1.11	0.90	1.09	0.92	1.29
Mean test (absolute value)	1.04	0.30	0.65	0.28	0.75

*, **, *** denote a significant test value on the 10%-, 5%- and 1%-level

The critical test values for the respective significance levels are as follows:

Chi^2 test (9 df): 14.68, 16.92, 21.67; Kuiper test: 1.62, 1.75, 2.00; Mean test: 1.64, 1.96, 2.58. They apply throughout the paper for any second digit analysis

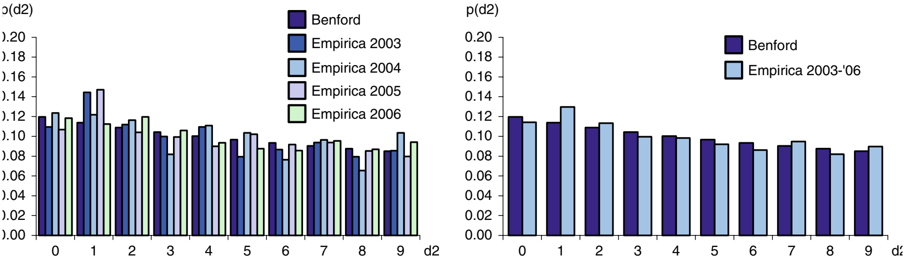


Fig. 2 Relative frequencies of the second digits of regression coefficients

At first glance, the figures for the years 2003, 2004 and 2005 do not appear to fit very well to the Benford distribution. But these discrepancies in the relative frequencies do not result in critical test statistics. However, as these deviations are still partly present in the total sample and given the higher number of observations the Chi^2 statistic falls into the rejection area.

If one suspects manipulation in the regression coefficients, our results indicate that first digits should be looked at. Intuitively, one would expect faked lower-order digits such that the regression outcomes support/refute a specific hypothesis. However, based on experimental evidence, Diekmann (2007) suggests to look at second-order digits.

3.3 Results for standard errors in Empirica

The same analysis is conducted for the standard errors. The results for the first digit are displayed in Table 4 and Figs. 3 and 4. The most interesting result is obtained for the year 2005 where all three test statistics reject the null on a 1% significance level.

Publication bias (Roberts and Stanley 2005) arises from the tendency of researchers, referees and editors to handle positive (and significant) results differently from negative (or insignificant) results. To chalk up a publication, authors may have an incentive, either through extensive data mining or—in the extreme case—by directly manipulating data or regression output, to turn

Table 4 Test statistics for the first digits of standard errors

Empirica	2003	2004	2005	2006	2003–06
Number of observations N	797	285	632	1,323	3,037
N per article	66	26	63	102	66
Chi^2 test	19.22**	9.99	29.33***	10.02	6.70
Probability	0.01	0.27	0.00	0.26	0.57
Kuiper test	1.45	0.76	2.46***	1.21	1.04
Mean test (absolute value)	0.33	0.13	2.71***	0.47	1.14

*, **, *** denote a significant test value on the 10%-, 5%- and 1%-level

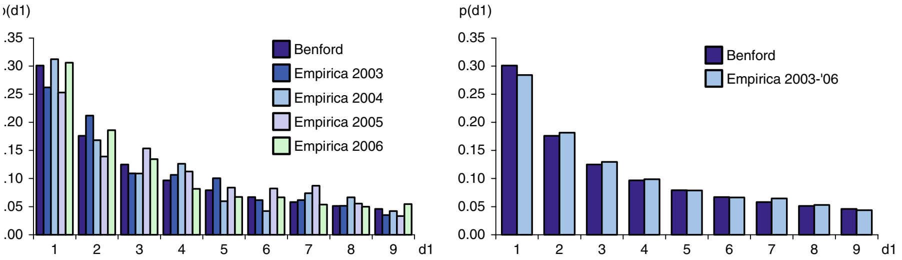


Fig. 3 Relative frequencies of the first digits of standard errors

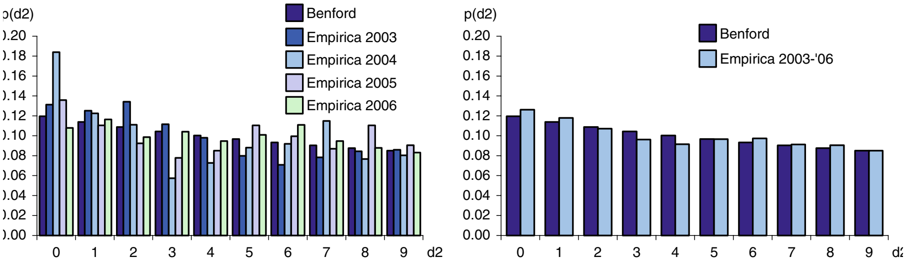


Fig. 4 Relative frequencies of the second digits of standard errors

insignificant key coefficients into significant ones. Thus, published regression results may be more likely to violate Benford's law, and, in particular, standard errors with t-values above 1.96 may be more likely having been engineered than those below 1.96. To check this issue, we divided the sample 2005 into two sub-samples which separates standard errors with t-values above 1.96 and below 1.96, respectively. It turns out that the dubious result is mainly caused by standard errors from the first sub-sample. In that region ($t > 1.96$, standard 5% significance level) the null hypothesis of a coefficient being zero is rejected. Therefore, one might tentatively argue that some statistics could have been amended in order to get significant regression coefficients. Another explanation might be that using the published (and rounded) data for calculations could yield to misleading results (see example in the beginning of this section). In 2005, only 38% of the 632 analysed

Table 5 Test statistics for the second digits of standard errors

Empirica	2003	2004	2005	2006	2003–06
Number of observations N	663	261	552	1,297	2,773
N per article	55	24	55	100	60
Chi^2 test	12.42	19.02**	12.51	8.09	6.05
Probability	0.19	0.03	0.19	0.52	0.73
Kuiper test	1.46	1.38	1.45	0.97	0.98
Mean test (absolute value)	2.18**	1.38	1.15	1.32	0.07

*, **, *** denote a significant test value on the 10%-, 5%- and 1%-level

standard errors were published, the remainder had to be calculated. For the whole sample (incl. 2005) more than half of the S.E.s were available without calculation. All other test statistics—with exception of the Chi^2 test value for 2003—indicate accordance of the first digits of standard errors with Benford's law.

The results for the second digit of standard errors show a high consistency with Benford's law. Only for 2003 the Mean test and for 2004 the Chi^2 test reject the null of a Benford distribution at a 5% significance level (see Table 5). It is worth noting that the second digits of the year 2005 do not exhibit any irregularities. Consequently, our arguments for possible manipulations above may be substantiated.

3.4 Results for coefficients and standard errors in Applied Economics Letters

The above findings shall be checked by analysing all articles published in *Applied Economics Letters* 2006. It can be seen from the test statistics displayed in Table 6 that regression coefficients (no graphics shown) exhibit a distribution approximately equal to Benford's law. In contrast, the test statistics (except the Mean test) for the first and second digits of standard errors are highly significant. This is also graphically illustrated (Fig. 5). Regarding the first digit, there is an excess of ones whereas a lack of nines for the second digit mainly causes the dubious statistics. Dividing the sample into two sub-samples of S.E. classified by the implied t-values (below or above $t = 1.96$), the results are ambiguous: The dubious test statistics for the first digit seem to be caused by the sub-sample with $t < 1.96$, whereas the reverse is true for the second digit. However, the problems might again be caused by rounding effects since approximately 63% of the first digits had to be calculated.

Again it can be shown that by only looking at digits that are not rounded the results deteriorate dramatically. Regarding the regression coefficients, the Chi^2 test and the Mean test reject the null of a Benford distribution at a minimum 5% significance level (the Kuiper test shows significant results only for the second digit). Interestingly, although the results for the standard errors get worse too, the Mean test still shows no significance on a 10% level.[6]

[6] For both journals also the possible sequences of the first and second digits (e.g. 14, 73, 86) have been analysed. The results, which are not reported here, show no clear pattern, neither regarding the tendencies of tests (which rejects more often) nor the effects of sample size.

Table 6 Test statistics for regression coefficients and standard errors

Applied Economics Letters 2006	Regr. coefficients		Standard errors	
	1st digit	2nd digit	1st digit	2nd digit
Number of observations N	5,171	4,650	2,921	2,619
N per article	73	65	41	37
Chi^2 test	7.23	14.27	48.01 ***	25.64 ***
Probability	0.51	0.11	0.00	0.00
Kuiper test	0.81	1.17	3.25 ***	1.66*
Mean test (absolute value)	0.31	1.77 *	1.25	0.89

*, **, *** denote a significant test value on the 10%-, 5%- and 1%-level

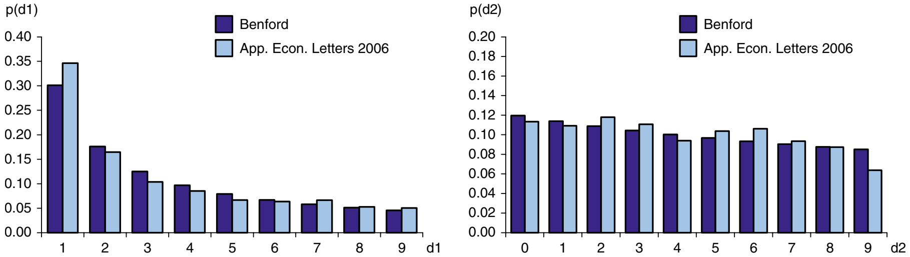


Fig. 5 Relative frequencies of the first and second digits of standard errors

Overall, the results suggest that in economic research Benford's law applies to regression coefficients and standard errors. Given the large sample sizes, the probabilities for a type II error, i.e. falsely accepting the null hypothesis of Benford's law, are very small. Nonetheless, in some cases there are doubts about the reliability of first digits but none for higher order digits. While the results for the regression coefficients are robust, the analysis (and consequently the interpretation) of the standard errors was restrained by limited data availability.

3.5 Problems with testing single articles

It is not the aim of this study to identify specific articles which might include irregularities in the regression results. Nevertheless, one can ask whether the tests would be able to detect manipulation if it were present, given the small sample size of first or second digits in a typical article. As shown in Table 2 (6), in *Empirica (Applied Economics Letters)* there were on average 100 (73) first digits per article. In general, a manipulation on digit $\mathrm{d_1}$ changes the ratio of that digit from $\mathrm{p_{d_1}}$ to $\mathrm{h_{d_1}} = \mathrm{p_{d_1}} + \delta_{\mathrm{d_1}}$, where the contamination ratios $\delta_{\mathrm{d_1}}$ are restricted such that the relative frequency of a certain digit remains between zero and one and $\sum_{\mathrm{d_1}=1}^{9} \delta_{\mathrm{d_1}} = 0$ holds. Whether a certain manipulation moves the test statistics into the critical region or not depends on the significance level (α), the sample size (N) and also on the pattern of deviations from Benford's law. For example, manipulation may change all digits (e.g. decreasing the relative frequency of leading digit 1 by some amount and increasing the frequencies of all other digits proportionally). Or only two digits may be affected (e.g. increasing the frequency of leading digit 5 at the expense of digit 1).

Given the sample size and the significance level, critical contamination ratios can be readily calculated for both types of contamination. For the first pattern of manipulation mentioned above, the Mean test performs best in the sense that it yields the smallest critical contamination ratio (9%, compared to 16% for the other two tests, at $N = 100$ and $\alpha = 10\%$). For the second pattern the Chi^2 test yields the smallest critical ratios (9%, compared to 10% for the Mean test and 16% for the Kuiper test). Thus, detecting fraud at conventional significance levels of 5 or 10 percent in a typical article with 100 regression coefficients requires fairly heavy manipulation. At the same time a probability of a type II error (ß) of around 37 percent for the Mean test is implied at the critical contamination ratios. Leamer (1978, p. 98) criticized the mechanical rule to *"set* $\alpha = 0.05$*"* regardless of the sample size in classical hypothesis testing. As a remedy, the significance level could be increased markedly in small samples, yielding a more balanced assignment of both types of error.

4 Benford's law in published economic forecasts

Monetary policy decisions by central banks on setting interest rates and by national governments on fiscal policies are informed by forecasts of macroeconomic variables. The growth rate of the real gross domestic product (GDP) and the inflation rate of the consumer price index (CPI) are undoubtedly at the centre of interest. Such forecasts stem from both, private and publicly funded institutions (e.g. investment banks and research institutes; in the following referred to as *institutes* or *panellists*). The *Consensus Forecasts* survey published by the London-based company *Consensus Economics* belongs to one of the broadest survey data sets available for macroeconomic research. The journal does not only report the mean forecasts of several macroeconomic variables for meanwhile more than 70 countries but also the data from each professional forecaster. The participating panellists are asked to provide their economic forecasts for the current and the subsequent calendar year on a monthly basis. Typically, forecasts are made by institutions located in the respective country of interest.

The *Consensus Forecasts* survey data are widely used and analysed in the literature. Batchelor (2001) finds that the *consensus* forecasts provided by *Consensus Economics* are more accurate and more informative than the forecasts of the International Monetary Fund and the World Bank for several macroeconomic variables of the G7 countries. Hendry and Clements (2004) outline theoretical reasons why generally *consensus* forecasts outperform single forecasts and support their analysis by Monte Carlo simulations. Isiklar and Lahiri (2007) use monthly GDP data from *Consensus Economics* for 18 developed countries and find that the predictive power of forecasts is low when the forecast horizon exceeds 18 months. However, only few studies make use of the disaggregated data of individual forecasters published in *Consensus Forecasts*. Harvey et al. (2001) analyse forecasts from several panellists for the United Kingdom GDP growth rate, unemployment rate and the growth rate of retail prices to assess forecast efficiency. Gallo et al. (2002) analyse data for the United States, the United Kingdom and Japan and find that forecasters have an imitation or herding behaviour and the tendency to converge to the mean forecast. This yields severe consequences, e.g. the standard deviation of the mean forecast can not be used as a valid measure of uncertainty. Dovern and Weisser (2007) analyse the forecasting accuracy of single panellists for four macroeconomic variables for the G7 countries. Osterloh (2008) investigates a wide range of consensus forecasts for Germany. However, none of the above mentioned studies takes the approach chosen in this paper.

In this study, we make use of the disaggregated data and investigate forecasts for the German real GDP growth rate and the inflation rate (measured as the change of the consumer price index, CPI). The data analysed run from October 1989 to July 2004. Specifically, we investigate if Benford’s law applies to the second digits of the single forecasts. In this context, the second digit is defined as the “first digit after decimal point”. Obviously, it is not plausible to check the first digit (before decimal point) for accordance with the Benford distribution since they are mostly in the range from zero to four for the GDP growth rate and inflation. For example, 46% of the first digits of the realised CPI growth rates start with the digit “1” and another 20% with the digit “2”[7], in sharp contrast to Benford’s law. Nevertheless, the second digits broadly conform to Benford’s law. However, for the “first digits after the decimal point” the agreement with Benford’s law is even better.[8] Moreover, the “first digits after the decimal point” better represent the economic weight of the digits with respect to rounding. For example, for the growth rates 1.25 and 0.92 the digit “2” carries not the same weight but would be counted as second digit according to Benford in both cases.

During the sample period some changes in the forecast variables published for Germany have occurred: Until December 1992, panellists had to report the gross national product for West-Germany. From January 1993 onwards this was replaced by the gross domestic product: At first only for West-Germany, but finally for the unified Germany (since May 1997). The shift in the CPI from West-Germany to the unified Germany took place in October 1996. In addition, the structure of panellists is not the same for the sample period: The number of participation institutes (around 25) varied across time and some institutes merged with others, were acquired or even went bankrupt. However, we neglect these effects in our analysis. Forecast values equal to 0.0 are included as well.

A specific feature of the data is that all published forecasts are restricted to one digit after the decimal point. This suggests that each participating panellist is forced to round its (possibly more precise) predictions before submitting it to the journal. Hence, it is necessary to adjust Benford’s law to take account for such rounding effects. The new distributions for the first and second digit are listed in Table 7. Suppose, one rounds to only one leading digit, then for example “3” as a first digit appears for all (not rounded) values between 2.5 and <3.5 (rounded: 3.00) with

[7] Based on an analysis of the realised CPI growth rates for Germany (10/1989–07/1994).

[8] The Chi^2 statistic for the second digits is 11.61 (p-value: 0.24) and for the “first digit after the decimal point” 5.61 (p-value: 0.82).

Table 7 Rounded Benford distribution

d	0	1	2	3	4	5	6	7	8	9	Mean
p(d1_rd)	n.a.	0.198	0.222	0.146	0.109	0.087	0.073	0.062	0.054	0.048	4.193
p(d2_rd)	0.103	0.117	0.111	0.107	0.102	0.098	0.095	0.092	0.089	0.086	4.761
p(d2_bold_rd)	0.506	*	*	*	*	0.494	*	*	*	*	
p(d2_mix)	0.222	0.082	0.078	0.075	0.072	0.216	0.067	0.065	0.063	0.061	4.231

Source: Own calculation

probability 0.146. Accordingly, if the second digits are rounded, for example “4” appears for all values between 1.35 and <1.45 (rounded: 1.4). The final column shows that rounding also distorts the mean of the distribution of first digits (from 3.940 to 4.193) and of second digits (from 4.687 to 4.761). The third row applies for the case that the second digits are boldly rounded to half-percentage points such that only zeros and fives are reported as second digits. The final row is a mixture of both, as will be explained below.

We start by presenting the results of four time series for the whole observation period and all panellists: The real GDP growth rate and the inflation rate for the current and the subsequent year. The test statistics are displayed in Table 8, a graphical illustration is given in Fig. 6.

There are more than 4,400 observations for each time series and 55 panellists in total. The exact number of observations varies across the series’ since not every panellist reports figures for all asked variables at each record date. As it can easily be seen, all test statistics (except one) are highly significant. Thus, the null of a (rounded) Benford distribution for the second digits has to be rejected. The graphics show that this is due to an excess of zeros and fives in the forecasts. This effect is similarly strong for all four time series and consequently for the pooled sample. In other words, in approximately 23% of all data, the forecasts look like 0.0 and in 21% like 0.5 (with any first digit).

A priori, the value added from asking many (instead of a few) professional forecasters for their opinions is higher forecast accuracy. Therefore, one would expect at least a difference in the forecasts of the individual institutes for the second digit (the first digit

Table 8 Test statistics for the second digits of consensus forecast data

Variable	GDP		CPI		Total
Forecast period (year)	Current	Subseq.	Current	Subseq.	Sample
Number of observations N	4,652	4,445	4,697	4,498	18,292
Chi^2 test	1,934 ***	2,434 ***	1,345 ***	1,632 ***	7,048 ***
Probability	0.00	0.00	0.00	0.00	0.00
Kuiper test	12.67 ***	13.21 ***	14.02 ***	10.58 ***	25.22 ***
Mean test (absolute value)	7.95 ***	10.63 ***	2.17 **	7.72 ***	11.97 ***

*, **, *** denote a significant test value on the 10%-, 5%- and 1%-level

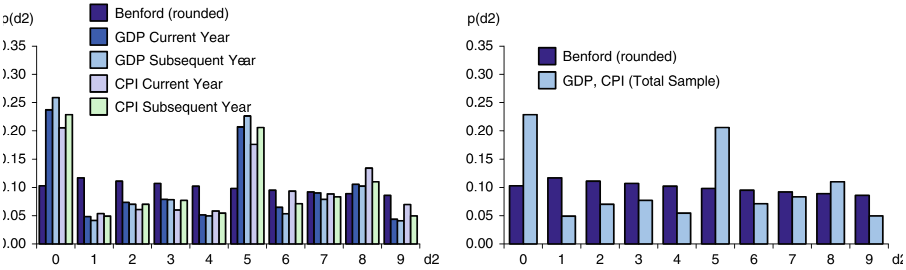


Fig. 6 Relative frequencies of the second digits of *Consensus Forecasts* data

should in most cases be the same across institutes for one prediction period). Figure 7 shows the distribution for the second digits of the pooled sample (i.e. all four time series) for each of the nine most mentioned institutes (*A, B,…, I*). The selection does not account for mergers and acquisitions of institutes. The sample size for each institute lies in the range of (roughly speaking) 600 and 700 observations which can be seen as sufficiently large (the sample size for the individual time series is too low to be analysed by institute). At first glance, institutes C, D and G graphically have quite a broad distribution. Nonetheless, even for these institutes zero and five are the most frequent digits (as it is the case for all others) and the accordance with the rounded Benford distribution is low. The worst outcome can be attributed to institute B, where in 70% of all data the second digit equals zero or five, and in another 18% equals eight, in sharp contrast to the rounded Benford distribution. Moreover, the second digits of the forecasts do not only deviate from the rounded Benford law but from the uniform distribution as well.

Given that this phenomenon (excess of zeros and fives) is present in all analysed samples above one might ask for the reasons. At first, one can think of model uncertainty: Suppose, an institute uses a model for prediction and the computation yields an inflation rate of, say 1.7361%. However, the forecaster knows that there is some uncertainty resulting from variables not incorporated in the model. To account for it, a rounding to 0.0 or 0.5 is done by a qualitative assessment of such factors. The mathematical consequences of such a clustering on zero and five are the following: Suppose, the population of forecasts obeys to Benford's law. If the rounding is such that all values which lie in the range of 0.75 and <0.25 are rounded to 0.0 and all other values are rounded to 0.5, than it can easily be checked by Monte Carlo simulations that the mean of the rounded data is biased (compared with the true mean). Imitation behaviour and information inefficieny (Osterloh 2008) may also play a role: if leading forecasters resort to bold rounding, other institutes might be inclined to follow. Furthermore, some institutes may report bold rounded figures because they reflect more or less educated guesses. In any case, it seems desirable to extent the existing forecasting methods.[9]

The foregoing results suggest that excess of zeros and fives in the distribution of second digits may be viewed as a mixture of proportion λ of forecasts with *"bold-rounded"* second digits and the proportion 1-λ of forecasts reported with two

[9] One promising approach has been proposed by Berlemann and Nelson (2005). They introduce a small-scale experimental stock market which yields the (mean) forecast of inflation rate as well as a likelihood measure for different inflation scenarios. The main idea is to use the market as the best instrument to uncover and aggregate private information.

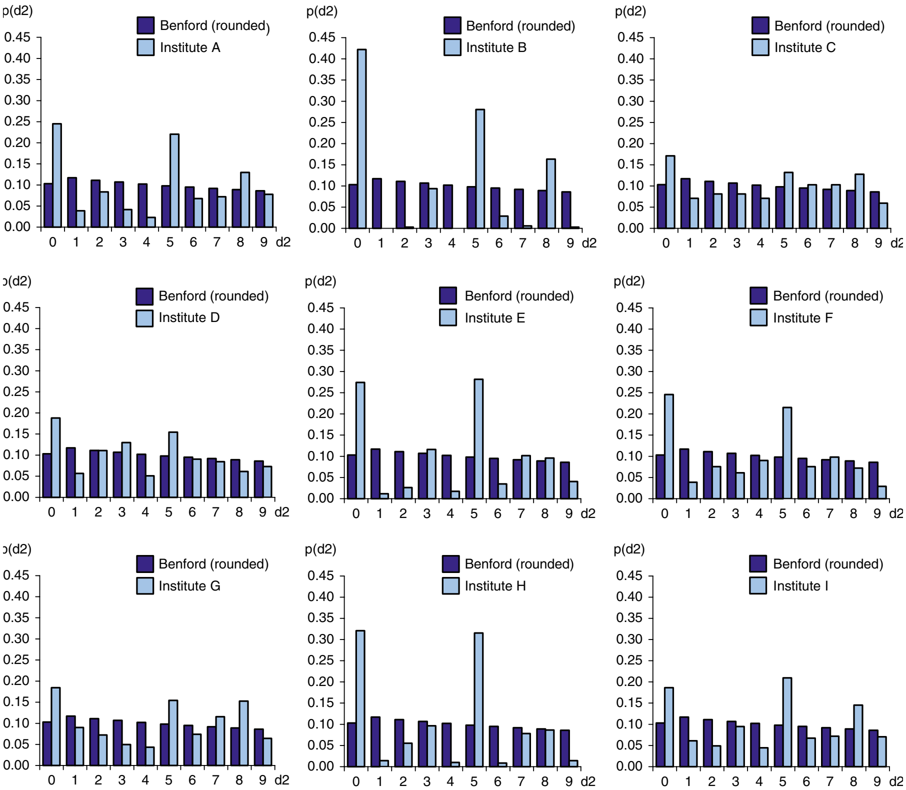


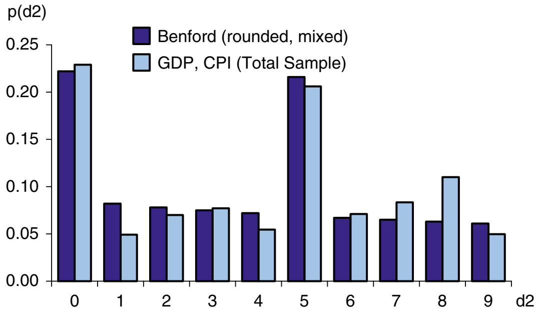
Fig. 7 Relative frequencies of the second digits of *Consensus Forecasts* data for various institutes

significant digits. Thus, the observed relative frequencies of second digits would have the mixed distribution

$$p(d_{2\_mix}) = \lambda\, p(d_{2\_bold\_rd}) + (1-\lambda)\, p(d_{2\_rd}) \tag{7}$$

with $0 \le \lambda \le 1$. Using the observed relative frequencies of second digits shown in Fig. 6, λ can be estimated by minimizing the sum of squared differences between $h(d_i)$ and $p(d_i)$ for i = 0…9. The estimated value is $\hat{\lambda} = 0.296$ (see Table 7). Thus, 30 percent of the observed forecasts seem to result from *“bold-rounding”* in the

Fig. 8 Mixed rounded Benford distribution vs. second digits of *Consensus Forecasts*



second digits, with little information content. Figure 8 shows the mixed rounded Benford distribution together with the distribution of the observed second digits of the total sample of *Consensus Forecasts*. The graphical fit is surprisingly good, in particular for the frequencies of zeros and fives, though there is an excess of digit eight and some deficit of ones.

5 Conclusions

This paper investigated the applicability of Benford's law in economic research and forecasting. We analysed the first and second digits of regression coefficients and standard errors in four volumes of *Empirica* and one volume of *Applied Economics Letters*, with almost 30,000 observations in total. In addition, we applied a rounded Benford distribution to the second digit of the GDP growth and CPI inflation rate forecasts for Germany drawn from 16 volumes of *Consensus Forecasts* with about 18,000 observations.

The main findings can be summarized as follows: Overall, published regression coefficients broadly conform to Benford's law. However, there are some irregularities with the first digit but none with higher-ordered digits. The results for standard errors do in general support Benford's law as well but are not that robust, possibly due to limitations of the available data. Thus, our results suggest that Benford's law can serve as a tool to assess the reliability of econometric research outcomes. Moreover, we found that checks for data manipulation should focus on the first digit. However, detecting deviations from Benford's law in single articles requires relatively high contamination ratios at conventional significance levels. The risk of overlooking doubtful papers can be reduced by increasing the significance level of the tests. In addition, it seems desirable that journal editors request from authors to report at least three significant digits of regression results and to provide both, standard errors and t-values.

In sharp contrast to regression coefficients, the second digits of economic growth and inflation forecasts exhibit a large excess of zeros and fives as the first digit after decimal point. Although the results vary slightly between different forecasters, they are very robust. An estimated share of 30 percent of the forecasts appears to be rounded to half percentage points, resulting in potentially severe information losses and, as a consequence, a distortion of the mean forecasts of the real growth rates and inflation rates.

Benford's law is a simple, objective and effective tool for detecting anomalies in large data sets that deserve closer inspection. Here, we looked at the output of economic research and forecasting. As Judge and Schechter (2007) observed, temptations for deception-prone activities may also be present in research input such as survey data.

References

Batchelor R (2001) How useful are the forecasts of intergovernmental agencies? The IMF and OECD versus the consensus. Appl Econom 33:225–235

Benford F (1938) The law of anomalous numbers. Proc Am Philos Soc 78:551–572

Berlemann M, Nelson F (2005) Forecasting inflation via experimental stock markets: some results from pilot markets. Ifo Working Paper No. 10

Camerer CF (2003) Behavioural game theory: experiments in strategic interaction. Russell Sage Foundation and Princeton University Press, New York, NY

Carslaw C (1988) Anomalies in income numbers: Evidence of goal oriented behaviour. Account Rev 63:321–327

De Ceuster MJK, Dhaene G, Schatteman T (1998) On the hypothesis of psychological barriers in stock markets and Benford's Law. J Empir Finance 5:263–267

Diekmann A (2007) Not the first digit! Using Benford's law to detect fraudulent scientific data. J Appl Stat 34:321–329

Dovern J, Weisser J (2007) Survey expectations in G7 countries: professional forecasts of macroeconomic variables from the consensus data set. The Kiel Institute for the World Economy, Mimeo

Durtschi C, Hillison W, Pacini C (2004) The effective use of Benford's law to assist in detecting fraud in accounting data. J Forensic Account 5:17–34

Gallo GM, Granger CWJ, Jeon Y (2002) Copycats and common swings: the impact of the use of forecasts in information sets. IMF Staff Pap 49:4–21

Giles DE (2007) Benford's law and naturally occurring prices in certain ebaY auctions. Appl Econ Lett 14:157–161

Hamermesh DS (2007) Viewpoint: replication in economics. Can J Econ 40(3):715–733

Harvey DI, Leybourne SJ, Newbold P (2001) Analysis of a panel of UK macroeconomic forecasts. Econom J 4:37–55

Hendry DF, Clements MP (2004) Pooling of forecasts. Econom J 7:1–31

Hill TP (1995) A statistical derivation of the significant-digit law. Stat Sci 10:354–363

Hill TP (1998) The first digit phenomenon. Am Sci 86:358–363

Isiklar G, Lahiri K (2007) How far ahead can we forecast? Evidence from cross-country surveys. Int J Forecast 23:167–187

Judge G, Schechter L (2007) Detecting problems in survey data using Benford's Law, November 1, Working Paper. University of California and University of Wisconsin

Kuiper NH (1959) Alternative proof of a theorem of Birnbaum and Pyke. Ann Math Statis 30:251–252

Leamer E (1978) Specification searches ad hoc inference with nonexperimental data. John Wiley & Sons, Inc., New York

Ley E (1996) On the peculiar distribution of the U.S. stock indexes' digits. Am Stat 50:311–313

McCullough BD, Vinod HD (2003) Verifying the solution from a nonlinear solver: a case study. Am Econ Rev 93:873–892

McCullough BD, McGeary KA, Harrison TD (2006) Lessons from the JMCB archive. J Money Credit Bank 38(4):1093–1107

Mochty L (2002) Die Aufdeckung von Manipulationen im Rechnungswesen–Was leistet das Benford's Law? Die Wirtschaftsprüfung 14:725–736

Newcomb S (1881) Note on the frequency of use of the different digits in natural numbers. Am J Math 4:39–40

Nigrini MJ (1996a) A taxpayer compliance application of Benford's law. J Am Taxpayer Assoc 18:72–91

Nigrini MJ (1996b) Using digital frequencies to detect fraud. The White Paper (April/May) 3–6

Nigrini MJ (1999) Adding value with digital analysis. Intern Auditor 56:21–23

Niskanen J, Keloharju M (2000) Earnings cosmetics in a tax-driven accounting environment: evidence from Finnish public firms. Eur Account Rev 9:443–452

Osterloh S (2008) Accuracy and properties of German business cycle forecasts. Appl Econ Q 54(1):27–57

Pinkham RS (1961) On the distribution of first significant digits. Ann Math Stat 32:1223–1230

Quick R, Wolz M (2003) Benford's law in deutschen Rechnungslegungsdaten. Betriebswirtschaftliche Forschung und Praxis 208–224

Reulecke A-K (2006) Fälschungen – Zu Autorschaft und Beweis in Wissenschaften und Künsten. Eine Einleitung. In Reulecke A-K, Fälschungen. Suhrkamp Verlag, Frankfurt am Main, pp 7–43

Roberts CJ, Stanley TD (2005) Meta-regression analysis: issues of publication bias in economics. Blackwell Publishing, Oxford, UK

Schatte P (1988) On mantissa distributions in computing and Benford's law. J Inf Process Cybern 24:443–455

Schäfer C, Schräpler JP, Müller KR, Wagner GG (2005) Automatic identification of faked and fraudulent interviews in the German SOEP. Schmollers Jahrbuch—J Appl Soc Sci Stud 125:119–129

Schräpler JP, Wagner GG (2005) Characteristics and impact of faked interviews in surveys. All Stat Arch 89:7–20

Tam Cho WK, Gaines BJ (2007) Braking the (Benford) law: statistical fraud detection in campaign finance. Am Stat 61(3):218–223

Thomas JK (1989) Unusual patterns in reported earnings. Account Rev 64:773–787

Tödter K-H (2007) Das Benford-Gesetz und die Anfangsziffern von Aktienkursen. Wirtschaftswissenschaftliches Studium 36(2):93–97

Van Caneghem T (2002) Earnings management induced by cognitive reference points. British Account Rev 34:167–178

Varian H (1972) Benford's law. Am Stat 23:65–66

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.